QTL analysis

VSNi

A Guide to QTL Analysis in Genstat® (22ⁿᵈ Edition)

by Martin Boer, Vanessa Cave, Hans Jansen, Marcos Malosetti,
Ky Mathews, Darren Murray, Fred van Eeuwijk and Sue Welham.

Genstat is developed by VSN International Ltd, in collaboration with practising statisticians at Rothamsted and other organisations in Britain, Australia, New Zealand and The Netherlands.

Contents

Introduction

QTL analysis is used to identify genetic factors underlying phenotypic variation in traits in a wide variety of contexts. The Genstat QTL system comprises a set of menus and commands to facilitate QTL analysis, bringing together a wide range of statistical techniques. The development has a particular focus on experimentation in plants but many of the techniques are more widely applicable. This Guide is designed to introduce you to these methods, and to enable you to use them correctly and effectively. The Guide focuses primarily on the menu interface, but also provides some information on use of commands.

Chapter 1 gives an overview of the Genstat QTL system, describes the different types of QTL analysis that can be done, the data required, and the different formats that can be used to import data into Genstat. The QTL Data Space can be used to manage these data structures and to capture the results of analysis. The QTL Data Space can be saved and reloaded to enable continuation of an analysis. A set of examples is introduced to demonstrate the range of analyses possible.

Chapter 2 gives details on how to import phenotypic and genotypic data structures into the QTL Data Space, and on the methods available to check, summarize and display the imported data structures.

QTL analysis in Genstat uses trait means for each genotype (or line), but usually data will be obtained from a replicated trial and so some preliminary analysis is required to obtain predicted means. Chapter 3 describes the facilities for such a preliminary analysis, and gives examples of some models commonly used for analysis of field trials.

Where several trials or experiments have been done in different environments, or at different times, it is often of interest to investigate differences in phenotypic response across environments, often termed genotype by environment (or G×E) interaction. Chapter 4 describes the use of a mixed model to quantify and describe the G×E interaction, and the use of other exploratory tools (AMMI and GGE biplots) to investigate the structure of the G×E interaction.

Chapters 5-8 describe QTL analysis for different populations of bi-parental lines. All of these models require that a genetic linkage map is available for the population. If that is not the case, Chapter 5 describes the facilities available for constructing and checking a map from a set of marker scores. Chapter 6 describes the statistical theory underpinning the Genstat QTL linkage models, including calculation of genetic predictors, marker

regression, simple interval mapping and composite interval mapping, with illustration for an inbred population with a single trait evaluated at a single site. Chapter 7 extends the analysis for inbred populations to the case where a single trait has been evaluated in multiple environments; or where multiple traits have been assessed on a single trial. Chapter 8 describes linkage analysis for cross-pollinated populations.

Chapter 9 describes linkage analysis for broader populations, often called association mapping analysis. This requires the consideration of population structure and linkage disequilibrium.

Chapter 10 gives some background information on linear mixed models, which form the analysis engine for the QTL system in Genstat. There is some technical detail alongside an example of a simple analysis of a field trial using the Genstat menu system.

Chapter 11 describes the use of commands for QTL analysis, which may be useful for very extensive data sets.

Acknowledgements: Paul Keizer and Jac Thissen for contributions to the QTL library.

1 Overview of the Genstat QTL system

QTL analysis is used to identify genetic factors underlying phenotypic variation in traits in a wide variety of contexts. The Genstat QTL system comprises a set of menus and commands to facilitate QTL analysis, bringing together a wide range of statistical techniques. Genstat's efficient algorithm for analysis of linear mixed models is used as the engine for QTL analysis. The QTL system has a particular focus on experimentation in plant populations but many of the techniques are more widely applicable, e.g. to animal or human populations.

This chapter describes:

- the different types of QTL detection possible in Genstat (Section 1.1)
- what data are required to perform QTL detection in Genstat (Section 1.1)
- the different types of experimental data that can be imported and processed (Section 1.1)
- how to use the QTL Data View panel (Section 1.2)
- the example data sets used in this Guide (Section 1.3)
- some analysis paths that can be used to detect and/or model QTLs (Section 1.4)

1.1   Methods for QTL detection in Genstat

Genstat's QTL system enables QTL detection for both structured populations and for more general association panels - we will describe the facilities and requirements for each of these in turn.

1.1.1   Linkage analysis

Linkage analysis can be performed for populations of inbred lines derived from an F1 cross of two homozygous parents as either F2 offspring, a back-cross of the F1 lines to one of the parents (BC1), double haploid offspring of the F1 generation (DH1), recombinant inbred lines of the nth generation (RILn), or back-crossed inbred lines (BCxSy). Linkage analysis is also possible for a cross-pollinated population derived from the cross of two heterozygous parents (CP). For all of these populations, linkage analysis can be done for a single trait in a single environment (Chapter 6) or across multiple environments, with estimation of QTL $\times$ environment (QTL$\times$E) interactions (Chapter 7). Alternatively, linkage analysis can be done for multiple traits in a single environment, allowing an effect of a QTL to be tested on multiple traits (Chapter 7).

These analyses require trait means for each line (or genotype) of the population (in each environment), together with genotype scores at a set of genetic markers and a genetic linkage map containing these markers. The set of genotype marker scores may contain missing values. Raw data from trials can be loaded and then subjected to a preliminary analysis to obtain trait means with standard errors (Chapter 3) which can be used as weights in linkage analysis (Chapters 6 and 7).

QTL estimation is done by simple or composite interval mapping using linear mixed models. Selection of candidate QTLs for use as cofactors can be automatic to manual, with back-selection used to obtain a final model. QTL scans can be plotted or saved. The selected QTLs can be saved with test statistics, confidence intervals and estimates of the QTL effects and standard errors.

1.1.2   Association analysis

Association analysis can be performed on broader populations for a single trait in a single environment or across multiple environments with estimation of QTL$\times$E interactions (Chapter 9). Identification of population substructure (i.e. genetic relatedness) is an essential step prior to this analysis, and can be achieved by explicit specification of population groups, by formation of a kinship matrix or by eigenanalysis (Patterson *et al.*,

2006) (Chapter 9). Modelling of the decay in linkage disequilibrium can be used to investigate the extent of linkage within the chromosomes.

These analyses require trait means for each line (or genotype) of the population (in each environment), together with genotype scores at a set of genetic markers and, to aid interpretation, a genetic linkage map containing these markers. The set of genotype marker scores may contain missing values. Rare alleles, with frequency below a user-specified threshold, may be excluded from analysis. Raw data from trials can be loaded and then subjected to a preliminary analysis to obtain trait means with standard errors (Chapter 3).

QTL estimation at each marker position, accounting for population substructure, uses a linear mixed model. The results of the association analysis at each of the marker positions can be saved as the significance level of the test, with the estimated effect and standard error for each allele.

1.2   The menu system and QTL Data View

The menus for QTL analysis can be found under the Stats menu on the menu bar (see Figure 1.1). They can also be accessed through the QTL Data View, which can be activated via the View QTL Data Space item (Figure 1.1) or under the View menu on the menu bar. The QTL Data View will then appear in a panel at the left-hand side of the screen (Figure 1.2). This space is shared with the Window Navigator and Data View (if activated) and the different views can be switched using the tabs at the bottom of the panel (Figure 1.2). The QTL Data View panel has three components.

- The top section provides a set of shortcuts for loading, deleting and saving data from the QTL Data Space, which provides a way of managing data structures to be used for QTL analysis.
- The middle section provides a shortcut to the menus found under the QTLs (Linkage/Association) menu item (Figure 1.1). The menu shortcuts can be displayed or removed by right-clicking on the QTL Data View and selecting the QTL Shortcut Menu item.
- The bottom section displays the data structures that have been loaded into the QTL Data Space.

Data structures can be added into the QTL Data Space when importing trait, map or marker data using the Data Import/Export menu within the QTL system (Chapter 2). Data structures within Genstat can also be added to the QTL Data Space using the icon at the

top of the QTL Data View panel ( ). Figure 1.3 shows the data space populated by trait means (under heading *Phenotypic means*), and map and marker data (under heading *Genotypic data*) prior to QTL analysis. The data structures in the QTL Data Space will be entered automatically into the QTL menus whenever possible. Where there is a choice of structures, such as a choice of phenotypic traits for analysis, by default only those in the QTL Data Space will be displayed as available data. Following QTL analysis, buttons on the top section of the QTL Data View pane can be used to save and export results to the Flapjack graphical genotyping tool (Milne *et al*., 2010) ( ) and to generate a report of the analysis ( ).
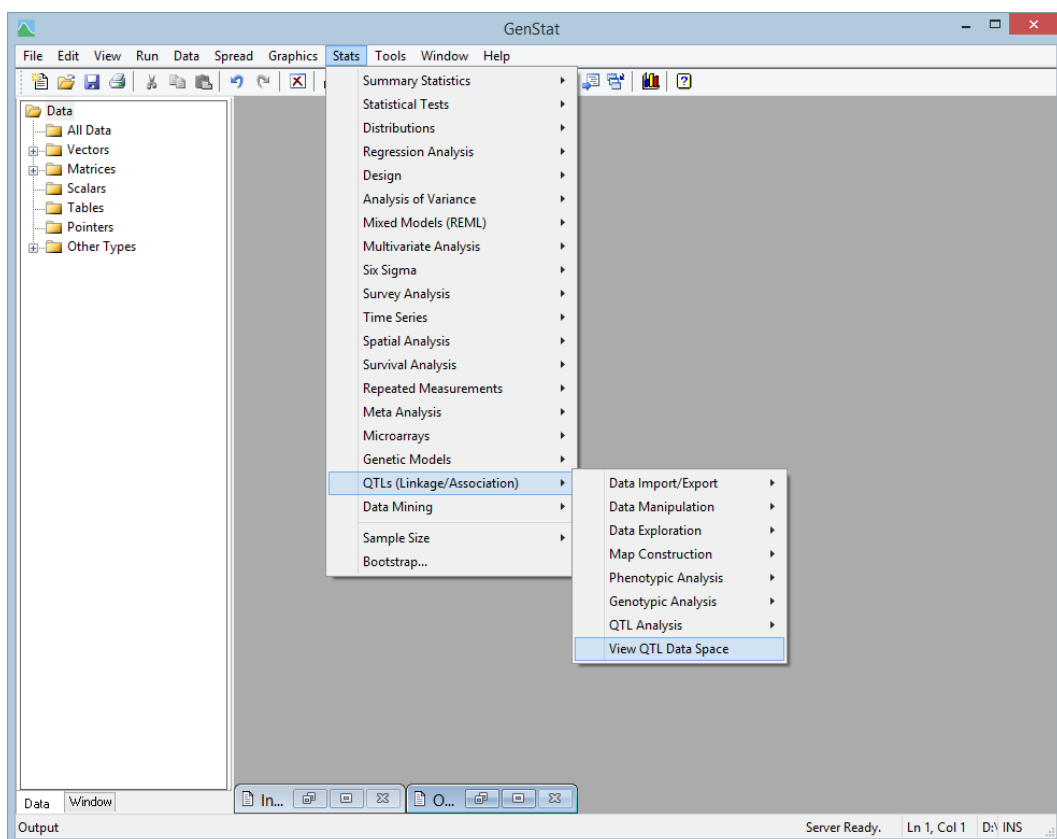
Figure 1.1: Accessing the QTL menu.

In this Guide, we show how to perform QTL detection using the menu system, but all of these analyses can be achieved using commands in the Genstat language. Use of the menus will generate commands in the Input Log window that can be used to repeat the analysis at a later date, if required. An overview of these commands is given in Chapter 11, and full documentation can be found within the Genstat Help System.
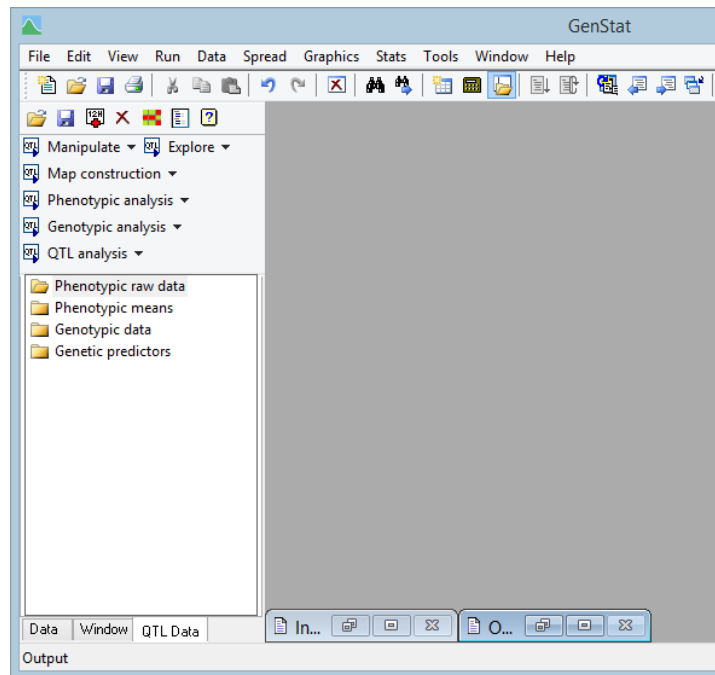
Figure 1.2: QTL Data View pane on left side of the Genstat window.
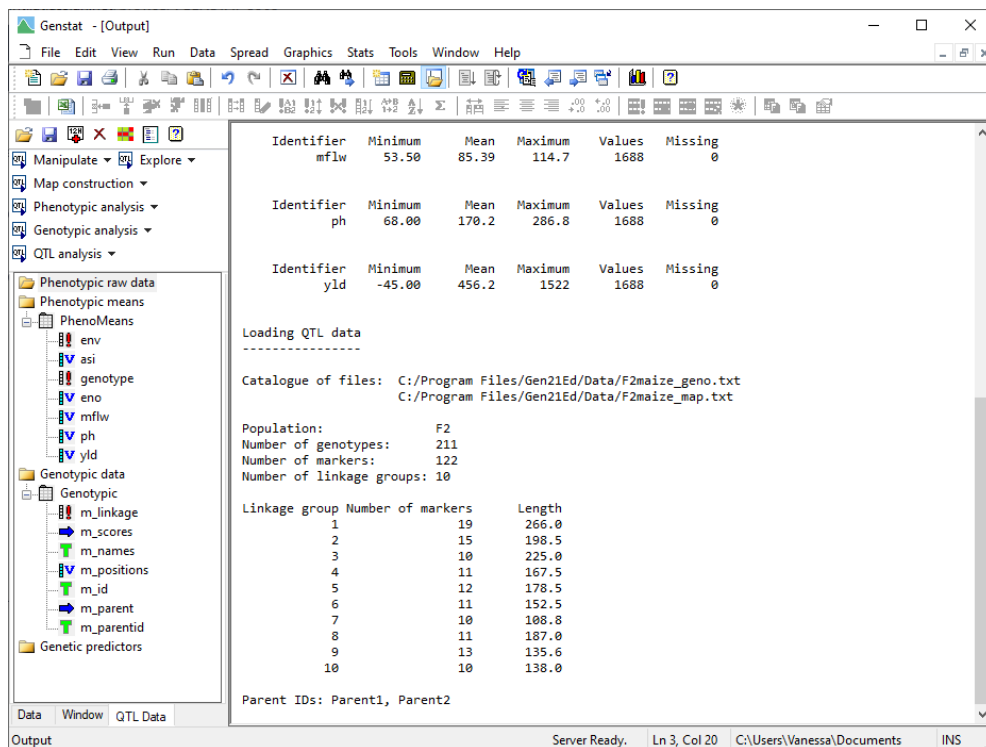
Figure 1.3: QTL Data View with trait means, marker scores and map information from the CIMMYT maize trials (Section 1.3.2) loaded prior to QTL analysis.

1.3   Example data sets

Within this Guide, we use several example data sets to illustrate the methods. The data files for these examples can be accessed via Help | Examples | Data Sets, then under Filter by topic: select A Guide to QTL Analysis. Alternatively, they can be found in `C:\Program Files\Gen21Ed\Data`. We introduce these data sets here in order to give some insight into the types of data that the Genstat QTL system is designed to deal with. In the next section, we indicate the types of analysis that might be appropriate for each data set.

1.3.1   Steptoe-Morex barley trial

This population is the well-known Steptoe $\times$ Morex double haploid population developed in the early 90s by the North American Barley Mapping Project. The objective was to improve the understanding of the genetic basis of agronomic and malting quality traits in barley. The population consists of 150 double haploid lines, which at that time was genotyped with 116 RFLP markers. The population was extensively evaluated for several agronomic and malting quality traits (Hayes *et al.*, 1993) in many locations and years (US and Canada). Here we use trait means for yield and heading date from one of those trials, held in file `SxM_pheno.csv`. The marker and map information are held in files `SxM_geno.txt` and `SxM_map.txt`, respectively.

1.3.2   CIMMYT maize trials

This data set comes from the maize drought stress breeding programme of the International Centre for Maize and Wheat Improvement (CIMMYT). The population is a F2 generated by crossing a drought tolerant parent (P1) with a drought susceptible one (P2). Seeds harvested from each of 211 F2 lines were used to test F3 families in 8 different environments; no, intermediate, and severe water stress trials in 1992 (`NS92a`, `IS92a`, and `SS92a` respectively), intermediate and severe water stress trials in 1994 (`IS94a`, `SS94a`), and low and high nitrogen in 1996 (`LN96a`, `LN96b`, and `HN96a`). The suffix 'a' indicates a winter trial, and 'b' a summer trial. The measured traits were: yield in kg/plot (`yld`), anthesis-silking interval in days (`asi`), number of ears per plant (`eno`), days to male flowering (`mflw`), and plant height in metres (`ph`). DNA was extracted from each of the 211 F2 plants to produce a total of 122 RFLP and AFLP markers covering the 10 maize chromosomes. Details of the data set can be found in the original publications (Ribaut *et al.*, 1996; Ribaut *et al.*, 1997). Trait means for each genotype from each trial are held in file `F2maize_pheno.csv`. The marker and map information are held in files `F2maize_geno.txt` and `F2maize_map.txt`, respectively.

1.3.3 CIMMYT spring wheat trials

This data set comprises raw plot data from a series of wheat trials conducted in Mexico by CIMMYT. The different trials took place under different regimes of irrigation and temperature, there were 4 trials across two years, labelled as `DRIP05`, `HEAT05`, `HEAT06`, `IRRI06`. Within each trial, a set of 167 progeny of a RIL (Recombinant Inbred Line; 8 generations) population were tested alongside the population parents (Seri and Babax). A lattice design with two replicates was used for each trial. In the first replicate the entries were not randomized, as they were considered to be a random selection from a population. At site `HEAT06`, the lattice was not exactly rectangular and so a check variety (`200`) was used to fill in the last row of the design. The yield for each plot at each site is given in file `SB_yield.csv`. The marker and map information are held in files `RILwheat_geno.txt` and `RILwheat_map.txt`, respectively; there are no marker scores for lines `SB004` and `SB084`.

1.3.4 MABDE barley association panel

A research programme (MABDE) was set up to investigate patterns of adaptation in barley. In this project a large set of barley genotypes (~190 genotypes) were evaluated in Europe and in the Mediterranean region. More details about this population and the research project can be found in Comadran *et al*. (2009). Here we look at yield in one of the environments, for a set of 179 genotypes. Mean yields for each genotype are held in `AMP_Barley_pheno.csv` with groups for association (linkage disequilibrium) mapping. The population was genotyped by DArTs. Marker scores and map information are held in files `AMP_Barley_geno.txt` and `AMP_Barley_map.txt`, respectively, and kinship data in `AMP_Barley_Kmatrix.txt`.

1.4 Example pipelines for QTL detection

In this section, we give an overview of the analysis path that we will follow for each of our example data sets in order to illustrate the different types of analysis that are available. Details of how to load the data and implement these analyses follow in subsequent chapters.

1.4.1 Steptoe-Morex barley trial

This data set comprises yield and heading date means from a single trial. The first step is to check these traits and the genotype marker scores and map for errors or inconsistencies (Chapter 2). We can then detect QTLs for each trait separately, using simple interval

mapping and/or composite interval mapping followed by model selection (Chapter 6). Finally, we can perform multi-trait QTL detection by joint modelling of the two traits (Chapter 7).
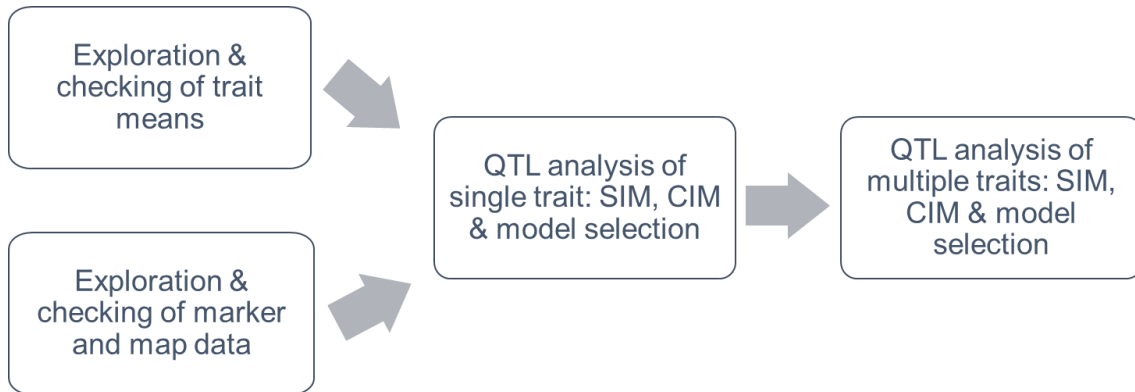
1.4.2 CIMMYT maize trials

This data set comprises means from 5 traits in 8 environments. The first step is to check these traits and the marker scores and map for errors or inconsistencies (Chapter 2). For each trait, we can then model the genotype by environment variation (G×E, Chapter 4) and use the results to investigate QTL by environment (QTL×E) interactions (Chapter 7).
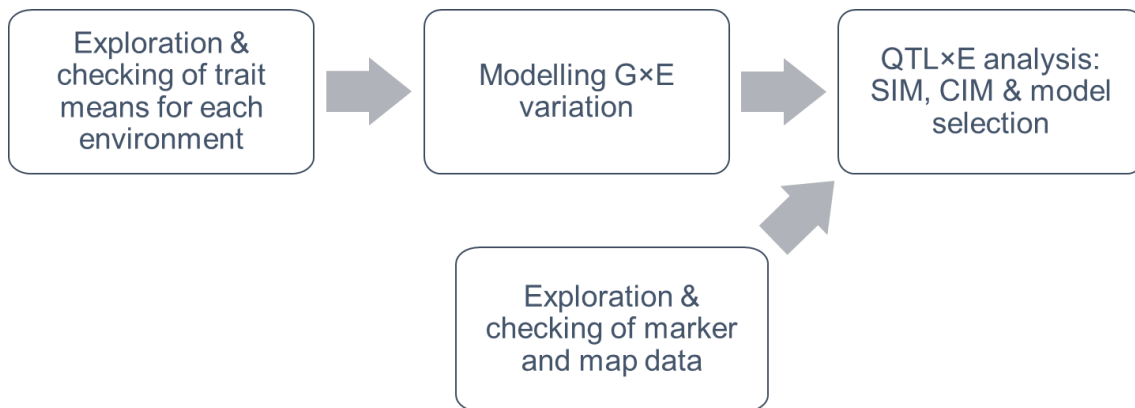
1.4.3 CIMMYT spring wheat trials

This data set comprises plot yields from 4 trials. The first step is to check the plot yields and the marker scores and map for errors or inconsistencies (Chapter 2). We then need to

analyse the raw yield data from each trial to obtain trait means for each genotype in each environment (Chapter 3). We can then model the genotype by environment variation (G×E, Chapter 4) and use the results to investigate QTL by environment (QTL×E) interactions (Chapter 7).
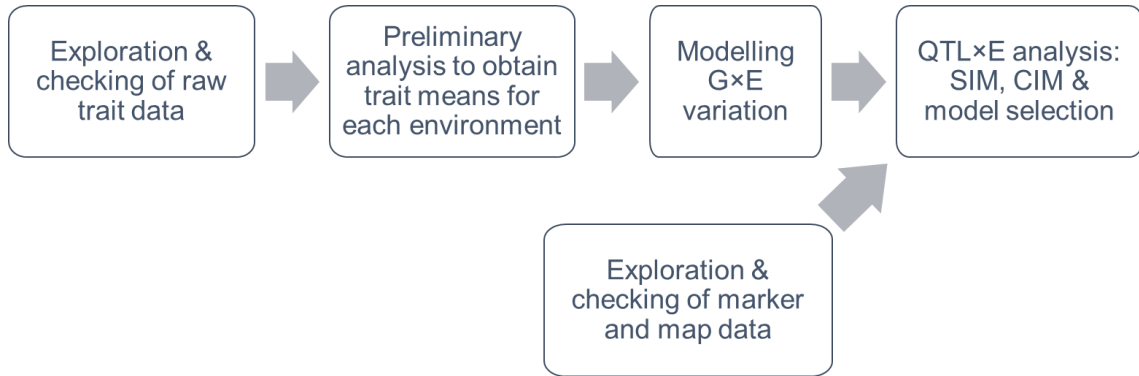
1.4.4 MABDE barley association panel

This data set comprises mean yields from a single trial. The first step is to check the yields and the marker scores and map for errors or inconsistencies (Chapter 2). Investigation of the population substructure is required before analysis, and can usually be captured by eigenanalysis (Chapter 9). Investigation of linkage disequilibrium can also give insight into the genetic structure of the population. Finally we can perform association analysis (Chapter 9).
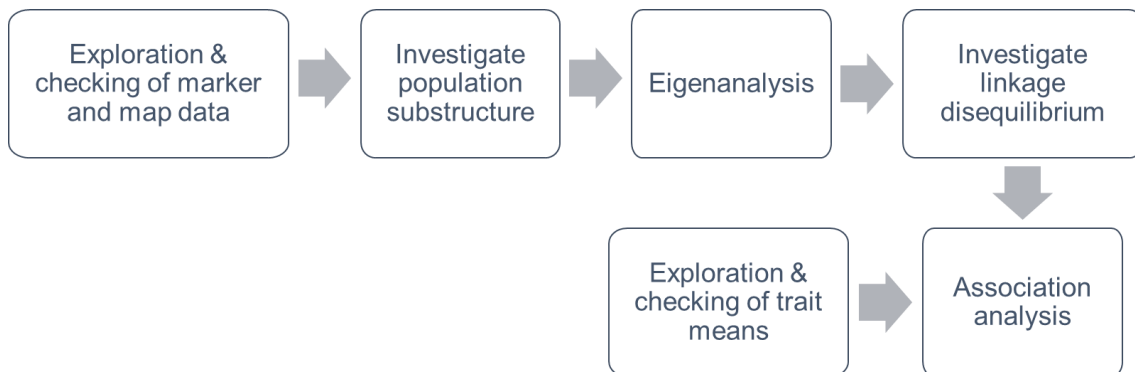
1.5   References

Comadran, J., Thomas, W.T.B., van Eeuwijk, F.A., Ceccarelli, S., Grando, S., Stanca, A.M., Pecchioni, N., Akar, T., Al-Yassin, A., Benbelkacem, A., Ouabbou, H., Bort, J., Romagosa, I., Hackett, C.A., & Russell, J.R. (2009). Patterns of genetic diversity and linkage disequilibrium in a highly structured *Hordeum vulgare* association-mapping population for the Mediterranean basin. *Theoretical and Applied Genetics*, **119**, 175-187.

Hayes, P.M., Liu, B.H., Knapp, S.J., Chen, F., Jones, B., Blake, T., Franckowiak, J., Rasmusson, D., Sorrells, M., Ullrich, S.E., Wesenberg, D., & Kleinhofs, A. (1993). Quantitative trait locus effects and environmental interaction in a sample of North American barley germ plasm. *Theoretical and Applied Genetics*, **87**, 392-401.

Milne, I., Shaw, P., Stephen, G., Bayer, M., Cardle, L., Thomas, W.T.B., Flavell, A. J., & Marshall, D. (2010). Flapjack - graphical genotype visualization. *Bioinformatics*, **26**, 3133-3134.

Patterson, N., Price, A.L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, **2**, 2074-2093.

Ribaut, J.-M., Hoisington, D.A., Deutsch, J.A., Jiang, C., & Gonzalez-de-Leon, D. (1996). Identification of quantitative trait loci under drought conditions in tropical maize. 1. Flowering parameters and the anthesis-silking interval. *Theoretical and Applied Genetics*, **92**, 905-914.

Ribaut, J.-M., Jiang, C., Gonzalez-de-Leon, D., Edmeades, G.O., & Hoisington, D.A. (1997). Identification of quantitative trait loci under drought conditions in tropical maize. 2. Yield components and marker-assisted selection strategies. *Theoretical and Applied Genetics*, **94**, 887-896.

2 Importing and checking phenotypic and genotypic data

Prior to any statistical analysis, it is necessary to import and check the data. For QTL analysis, both phenotypic (trait) and genotypic (marker and map) data are required. In this chapter we describe how to complete these tasks successfully using the QTLs (Linkage/Association) menu, which will load the imported data into the QTL Data Space (Section 1.2), ready for further analysis.

In this chapter you will learn how to:

- import phenotypic data, both raw data and pre-processed trait means (Section 2.1)
- import genotypic data, including marker, map and kinship data (Section 2.1)
- load and save a QTL Data Space (Section 2.2)
- manipulate data in the QTL Data Space (Section 2.3)
- summarize phenotypic and genotypic data (Section 2.4)

2.1   Importing data

Two types of data are required for a QTL analysis: phenotypic data (measurements of trait values) and genotypic data (evaluation of genotyping at markers and positions of markers on a genetic map). Following, we describe how to import phenotypic and genotypic data structures into the QTL Data Space.

2.1.1   Phenotypic data

Phenotypic data sets contain the quantitative traits (phenotypes) measured for all individuals (i.e. genotypes) in the population. Phenotypic data may be loaded as unprocessed data, i.e. raw measurements from a field trial or other experiment, or as pre-processed trait means for each genotype. We will consider the cases of raw data and trait means separately.

2.1.1.1   Raw data

Raw plot or unit data from an experiment (or several experiments) should consist of one or more columns of trait (phenotypic) measurements, a column specifying the genotype for each measurement and columns that identify the experiment and its structure, e.g. the trial name, and the origin of measurements from blocks and plots in the experimental design. This data file must have one row for each observation plus a header row indicating the column names. The data imported should comprise the whole experiment so that a valid analysis can be carried out; if only a subset of the genotypes are required for QTL analysis, this subsetting can be implemented as part of that analysis (see Section 2.3.2).

For example, Figure 2.1 shows the format of raw data from the CIMMYT spring wheat trials (Section 1.3.3) held in file `SB_yield.csv`. The trial (`Env`) and genotype (`Genotype`) labels for each measurement are given in the first two columns. These are coded as Genstat factors, indicated by using `!` at the end of the column name. The third-seventh columns give information on the trial design and field layout, to be used in estimation of trait means (Chapter 3), and the final column gives individual plot measurements of `yield`. Data in the same format can also be loaded from Genstat spreadsheets or workbooks, and Excel or tab-delimited text files.

Env!	Genotype!	Plot!	Rep!	Subblock!	Row!	Column!	yield
DRIP05	SB001	1	1	1	1	1	363
DRIP05	SB002	2	1	1	1	2	343
DRIP05	SB003	3	1	1	1	3	373
DRIP05	SB004	4	1	1	1	4	396
DRIP05	SB005	5	1	1	1	5	335
DRIP05	SB006	6	1	1	1	6	396
DRIP05	SB007	7	1	1	1	7	421
DRIP05	SB008	8	1	1	1	8	384
DRIP05	SB009	9	1	1	1	9	365
DRIP05	SB010	10	1	1	1	10	337
...	...	...	...	...	...	...	...

Figure 2.1: Raw phenotypic data from the CIMMYT spring wheat trials (Section 1.3.3) held in file `SB_yield.csv`: one row for each plot in each trial.

Raw phenotypic data can be imported either by clicking on menu items Stats | QTLs (Linkage/Association) | Data Import/Export | Load Phenotypic data (Figure 2.2) or by using the Open QTL data file icon (📂) on the QTL Data View and selecting the option Phenotypic data file. Both routes launch the Open Phenotypic Data files window (see Figure 2.3). Select the correct data type, with the setting Plot or unit data, then click on the Open button to initiate the Read QTL phenotypic plot data window (see Figure 2.3).
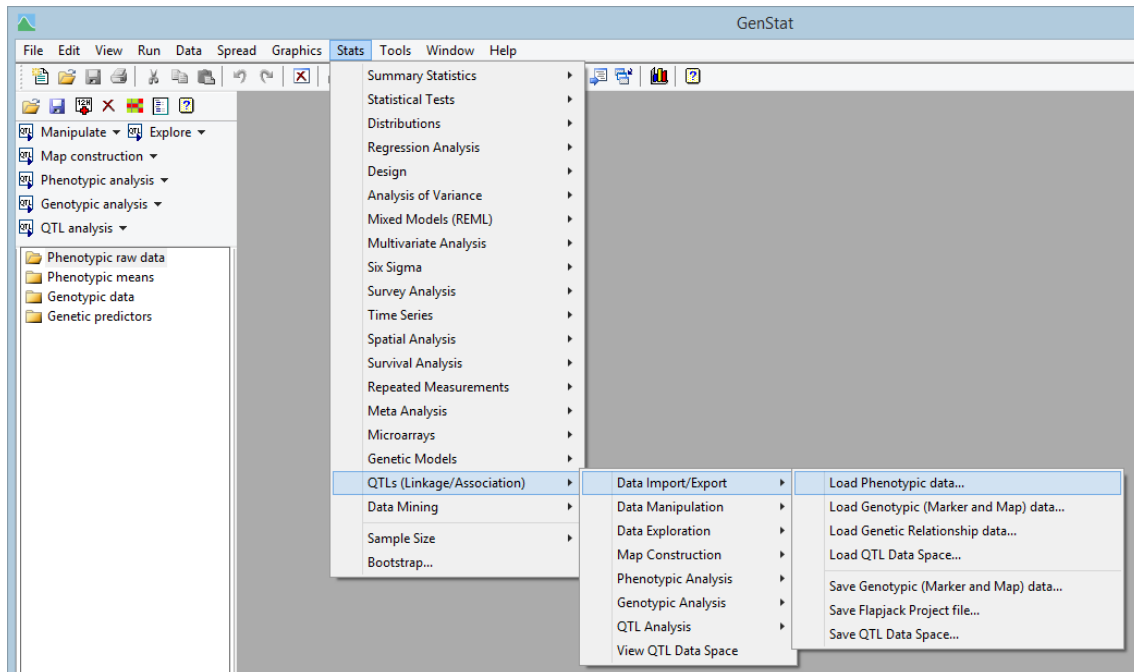
Figure 2.2: Accessing the Open Phenotypic Data files menu.
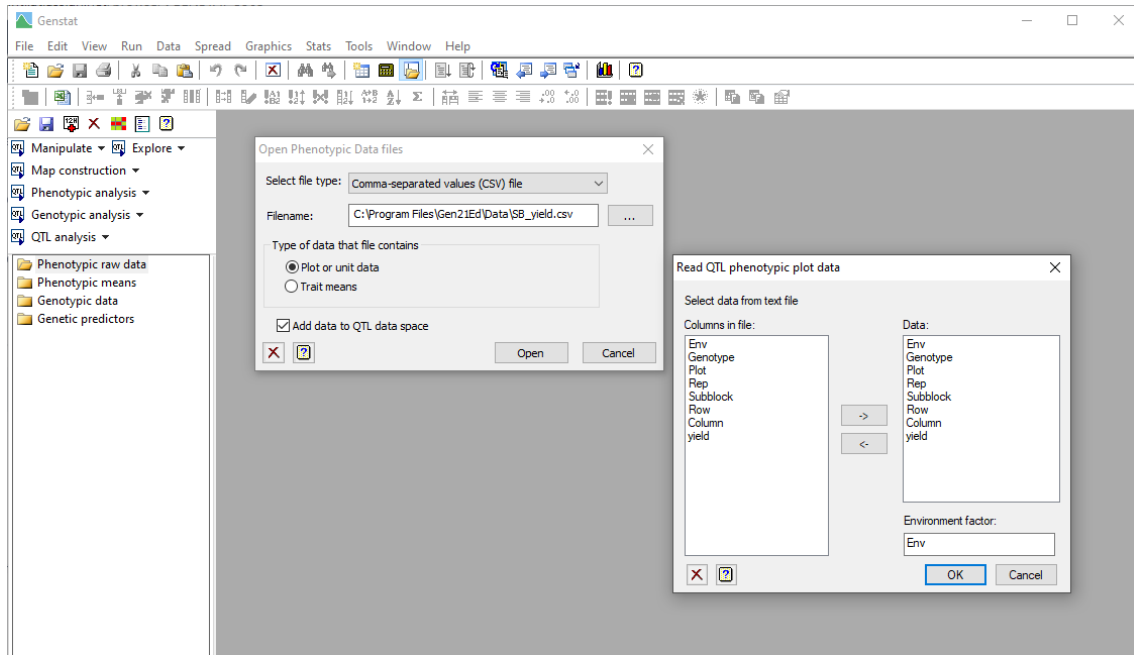
Figure 2.3: Importing phenotypic raw data from file `SB_yield.csv`.

All data that will be used in the analysis should be copied into the Data: box, and if the data arises from multiple experiments (environments), then the factor that labels these (here `Env`) should be entered as the Environment factor:; this factor will be used to subset the data for single trial analysis (Chapter 3). By default, the raw data will be added into the QTL Data Space (under *Phenotypic raw data*) and should appear in the QTL Data View (Figure 2.4) once the OK button has been pressed. The Output window will also show summaries of the variables imported (Figure 2.4).

Alternatively, raw data and associated factors can be loaded directly into Genstat (e.g. using File | Open) and then transferred into the QTL Data Space either by:

- selecting the Add data in Genstat to QTL data space icon (⏱) on the QTL Data View, followed by the Phenotypic raw data tab; or,
- right clicking on the *Phenotypic raw data* folder in the QTL Data View, then selecting Add to data space.
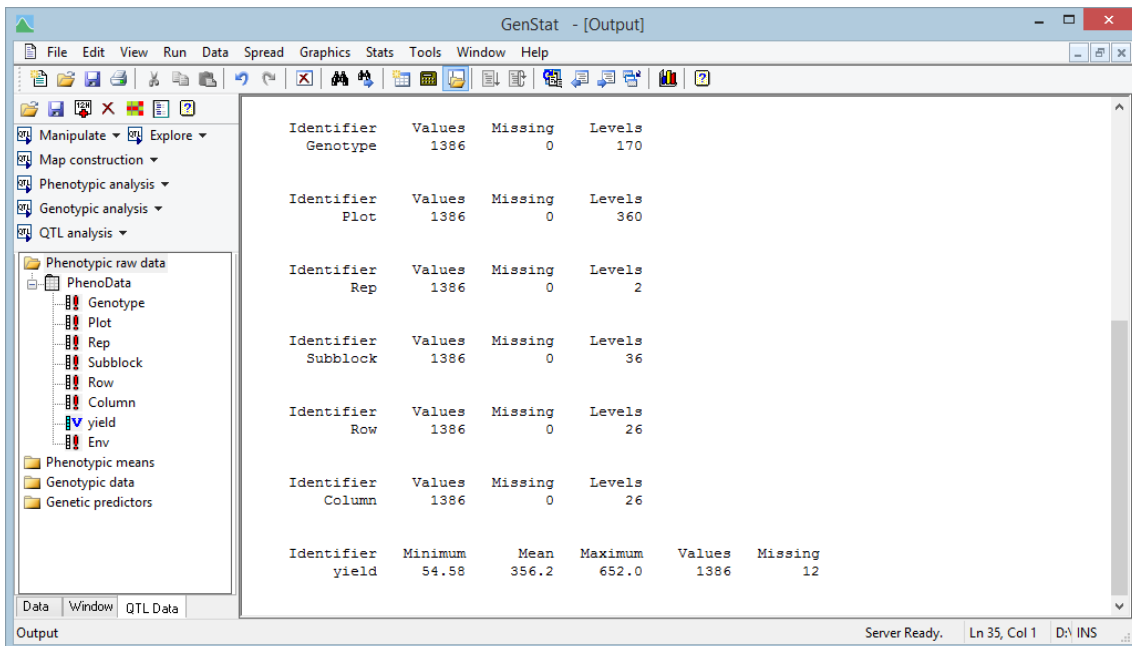
Figure 2.4: QTL Data View and Output window after importing raw data from file `SB_yield.csv`.

2.1.1.2 Pre-processed trait means

Trait means for each genotype are assumed to have been produced from an appropriate analysis of experimental data and ready for QTL analysis. For a single trial, this data file must have one row for each genotype plus a header row indicating the column names. If data from multiple experiments (environments) are provided, there should be one row for each genotype in each environment. The trait means must be classified by a factor indicating the individual genotypes and experiments and there must be only one value for each genotype × environment combination.

For example, Figure 2.5 shows the format of the trait mean data from the CIMMYT maize trials (Section 1.3.2) held in file `F2maize_pheno.csv`. The trial (`env`) and genotype (`genotype`) names for each measurement are given in the first two columns, followed by predicted means for five phenotypic traits: `asi` (anthesis-silking interval), `eno` (number of ears), `mflw` (days to male flowering), `ph` (plant height) and `yld` (yield). Data in the same format can also be loaded from Genstat spreadsheets or workbooks, and Excel or tab-delimited text files. Data in the .qua format of MapQTL® (van Ooijen, 2009) can also be loaded.

17

```
env!  genotype!      asi       eno      mflw        ph       yld
IS94a      G001      2.65      8.85     91.55     154.8     337.3
SS94a      G001      2.01      8.77     89.39     163.8     447.6
HN96b      G001       0.1      12.5      56.2       191       657
LN96b      G001       3.1       7.9      61.6       107        71
LN96a      G001       4.8       8.9      97.8        97       145
IS92a      G001     -2.35     11.58     92.94     205.4       672
NS92a      G001     -0.17     16.55     89.13     239.7      1260
SS92a      G001      1.07     10.21     90.14     205.8     493.4
IS94a      G002      2.44      9.92     88.43       173     603.1
SS94a      G002      3.43      8.53     84.63       167     331.5
  ...        ...      ...       ...       ...       ...       ...
```

Figure 2.5: Trait means from the CIMMYT maize trials (Section 1.3.2) held in file `F2_maize_pheno.csv`: one row for each genotype in each environment.

Phenotypic data can be imported using the Open Phenotypic Data files window, accessed either by Stats | QTLs (Linkage/Association) | Data Import/Export | Load Phenotypic data (Figure 2.2) or using the Open QTL data file icon (📂) on the QTL Data View and selecting the option Phenotypic data file. Figure 2.6 shows the Read QTL Phenotypic Means window, which is initiated from the Open Phenotypic Data files window when data type Trait means is specified (see Figure 2.3). The five traits have been copied into the Trait means: box, and the factors `genotype` and `env` have been assigned as the Genotype factor: and Environment factor:, respectively. When loaded, by pressing the OK button, these structures will appear in the QTL Data View in the *Phenotypic means* folder under category *PhenoMeans*.
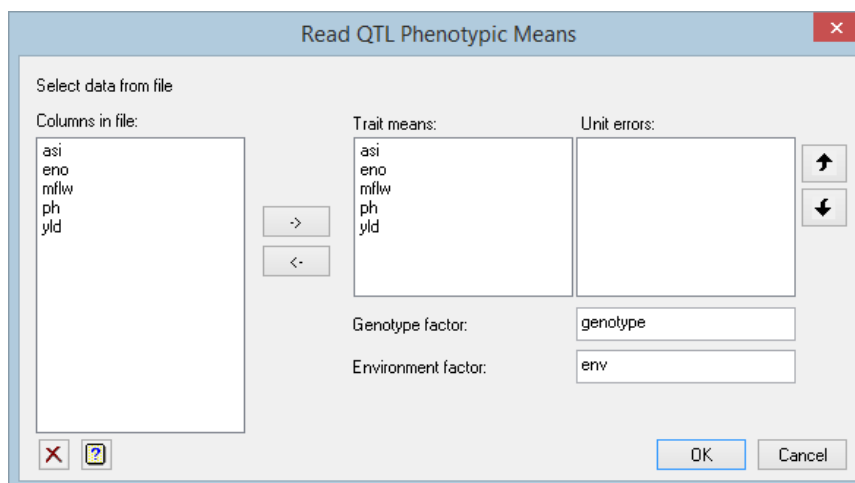
Figure 2.6: Window for importing trait means from the CIMMYT maize trials (Figure 2.5).

Alternatively, trait means can be loaded directly into Genstat (e.g. using File | Open) and then transferred into the QTL Data Space by either:

- selecting the Add data in Genstat to QTL data space icon (⊞) on the QTL Data View, followed by the Phenotypic means tab; or,
- by right clicking on the *Phenotypic means* folder in the QTL Data View, then selecting Add to data space.

2.1.2 Genotypic data

Genotypic data sets contain scores for a set of markers on each genotype and a genetic map containing those markers. Genstat can import genotypic data held in Flapjack text, R/qtl csvs or csvsr, or MapQTL® .loc and .map formats. Flapjack and MapQTL® formats both have two input files: a genotype file that contains a genotype by marker (locus) matrix of marker scores and a map file that contains the linkage groups and marker positions (using Haldane's distance in cM). The R/qtl formats consist of a single file containing both marker scores and map information. We give more details in the following subsections.

2.1.2.1 Flapjack marker and map files

The Flapjack (Milne *et al*., 2010) text file format is recommended as being the most flexible, allowing the user's choice of marker categories and providing some data checking/validation. This format can be used for any type of population.

The Flapjack format genotype file is a text file (tab delimited by default) containing a data matrix where rows correspond to parental lines (if present) and genotypes, and columns correspond to markers. The first row has an empty cell in the first column, followed by marker names. The second (and following) row(s) consist of the line name, followed by marker scores. For populations derived from one (or more) parental crosses, the parents must be listed before offspring. For these parental lines, a code is used to define the score for each parent at each marker. This code does not have to be a single character. For the offspring lines, codes for both parental alleles should be specified, separated by a separator (by default /), except for homozygotes, where a single code can be used. Note, in the case of association mapping populations, there are no parents.

For example, Figure 2.7 shows a portion of a marker score file (`F2maize_geno.txt`) in Flapjack format. This is an F2 population derived from two homozygous parents, so it is straightforward here to code the first parent as 1 and the second as 2 (we could have alternatively used A and B). Using these parental genotypes, the permitted codes for fully

informative markers on offspring are then: 1/1 (or 1), 1/2, 2/2 (or 2). Missing values are indicated by - (synonymous to -/-). In the case of partially informative markers (e.g. dominant markers) genotypes are coded as 1/- or 2/-, depending on whether the dominant allele originated from parent 1 or parent 2.

```
           L008   L058   L094   L040   L112   L062   L065   L085   L039   L082   L123   L070   L117   …
Parent1      1      1      1      1      1      1      1      1      1      1      1      1      1    …
Parent2      2      2      2      2      2      2      2      2      2      2      2      2      2    …
   G001    1/2    2/-    1/2    1/2    1/2    1/2    2/-    1/2    1/2    1/2    1/2      -    2/-   …
   G002    1/2    2/-    1/2    1/2      2      2    2/-      2      2    1/2    1/2      1      1   …
   G003    1/2    2/-      2      2      2      2    2/-      2      2      2      2      2    2/-   …
   G004    1/2    2/-      2      2      2      2    2/-    1/2    1/2    1/2    1/2    1/2    2/-   …
   G005      2    2/-      2      2      2      2    2/-      2      2      2      2      2    2/-   …
   G006      2    2/-    1/2    1/2    1/2    1/2    2/-      1      1      1      1    1/2    2/-   …
   G007    1/2    2/-      2      2      2      2      1      1      1    1/2    1/2    1/2    2/-   …
   G008      1      1      1      1      1      1    2/-    1/2    1/2      1      1      1    2/-   …
    …        …      …      …      …      …      …      …      …      …      …      …      …      …    …
```

Figure 2.7: Top section of Flapjack text file `F2maize_geno.txt`, which contains marker scores for the F2 population used in the CIMMYT maize trials.

The Flapjack format map file is a text file (again tab delimited by default) containing three columns which define the genetic map (see Figure 2.8). There is no header line. The first column contains the marker names, the second codes the linkage group that contains the marker, and the third column specifies the marker position within that linkage group.

```
                    L008        1        0
                    L058        1      7.4
                    L094        1     41.3
                    L040        1     73.4
                    L112        1     83.4
                    L062        1     84.8
                    L065        1    127.5
                    L085        1      141
                    L039        1    150.2
                    L082        1    174.8
                    L123        1    195.8
                    L070        1      206
                    L117        1    215.3
                     …          …       …
```

Figure 2.8: Top section of Flapjack text file `F2maize_map.txt` containing genetic map information for the F2 population used in the CIMMYT maize trials.

2.1.2.2   R/qtl csvs and csvsr files

R/qtl csvs and csvsr files (Broman and Sen, 2009) both use comma-separated format. The R/qtl csvs format is similar to the Flapjack format, but does not allow specification of parental codes and requires the map to be included in the same file. This format can only be used for F2, DH1, BC1, or RIL*n* populations.

The R/qtl csvs format genotype file is a .csv file containing a data matrix where rows correspond to genotypes, and columns correspond to markers. The first row has an identifier in the first column (used for the set of genotype labels), followed by marker names. The second row has an empty cell in the first column, followed by a label for the linkage group containing each marker. The third row has an empty cell in the first column, followed by the position of each marker within the linkage group. The fourth (and following) row(s) consist of the line name, followed by marker scores. Marker scores are coded using the characters A (homozygous like parent A), B (homozygous like parent B), H (heterozygous), C (not A) or D (not B). The default missing data code is '-'. For example, Figure 2.9 shows the top few rows of a marker score file (`F2maize_geno_csvs.csv`) in R/qtl csvs format.

```
  id   L008   L058   L094   L040   L112   L062   L065   L085   L039   L082   L123   L070   L117   …
         1      1      1      1      1      1      1      1      1      1      1      1      1    …
         0    7.4   41.3   73.4   83.4   84.8  127.5          141  150.2  174.8  195.8    206  215.3   …
 G00A    H      C      H      H      H      H      C      H      H      H      H      -      C    …
 G00B    H      C      H      H      B      B      C      B      B      H      H      A      A    …
 G003    H      C      B      B      B      B      C      B      B      B      B      B      C    …
 G004    H      C      B      B      B      B      C      H      H      H      H      H      C    …
 G005    B      C      B      B      B      B      C      B      B      B      B      B      C    …
 G006    B      C      H      H      H      H      C      A      A      A      A      H      C    …
 G007    H      C      B      B      B      B      A      A      A      H      H      H      C    …
 G008    A      A      A      A      A      A      C      H      H      A      A      A      C    …
  …      …      …      …      …      …      …      …      …      …      …      …      …      …    …
```

Figure 2.9: Top section of R/qtl csvs format combined marker and map file for the F2 population used in the CIMMYT maize trials.

The R/qtl csvsr format is a transpose of the csvs format, containing a data matrix where rows correspond to markers and columns correspond to genotypes.

2.1.2.3   MapQTL .loc and .map files

The MapQTL® format (van Ooijen, 2009) uses separate text files for the marker and map information. This format can be used for F2, DH1, BC1, RIL*n* or CP populations. Details of the permitted codes for each type of population can be found in van Ooijen (2009).

The top of the marker file for the CIMMYT maize F2 population (`F2maize.loc`) is shown in Figure 2.10. The first four lines of the marker file specify the population name (`name`) and type (`popt`), the number of markers (`nloc`) and the number of individuals (`nind`). These lines are followed by a list of scores for each marker. The number of markers and number of individuals must match that given at the start of the file. In order that Genstat can combine the marker and trait data without error, it is recommended that labels for the individuals are also given at the end of the .loc file (as shown in Figure 2.11). The order of the labels should exactly match the order in which the marker scores are listed.

```
name = F2maize
popt = F2
nloc = 122
nind = 211

L008  H H H H B B H A B H A H H H H H H H A - A H H B H H H H A H
H B H B A H B A B H H H H H B H H H H A B A H A A A H H B H H H A
H B H H H A A H H B B A B B A H A B B A H B H A H H B H A H A A H
A A H B H B A A H H H B B A H H B H H B A H A H H A A H H H H H B
H B H B B A B H B A A B A H B A A H B A B B H B A A H B A H H H H
B H B B H A A B B H B B B B A H H H H A A H A H H H A A H H H H A
H B H A A H A B A A A H B H H H
L058  C C C C C C C A C C A C C C C C C C A C A C C C C C C A C C
C C C C A C C A C C C C C C C C C C C C C A C A C A A A C C C C C A
C C C C C C A C C C C A C C A C A C C A C C C A C C C C A C A A C
A A - C C C A A C C C C C A C C C C C C A C A C - A A C C A C C C
C C C C C - C C C A A C C C C A A C C - C C C C C C C C A C C C C
C - C C C - C C C C C C C C A C C A C A C C A C C C A C C C - C -
C C C C A C A C A A A A C C A - C
L094  H H B B B H B A H H H H A H H H H H A H A H B H H H B H H B
H H H B A B B A H H B A - A B B H A H H B B B A A H H B H H H H -
H H B H B H H A H B B - B B A A H H H A H B H A B B B H H H H H H
B H H B H B H H H B H B H H B H B A H B A B A H B H H H H A H H H
B H B B B A B - B A A H H H B H A H H A B H A H B H H H H H H A H
B H A B H A B B H B B H H H H H H A H A H A B H H A H H H A H B A A H
H H H H H H H H H H H H H H A H B
...
```

Figure 2.10: Top section of the MapQTL file `F2maize.loc`, containing genotype scores at a set of genetic markers for the F2 population used in the CIMMYT maize trials.

```
L081  B B A H H H B A A H H B A B B H B A H H A H B H A H H A A B
H A A H H A H H A B H H - H H B H H H H H B A H H A A B A A H H -
A H A H H H A A B H B - A H H A B A B H A A H H H B H H H H A H H
H H B H H B H H A B H H A B H B B A A H A H H H H H B H H B B H
H H A B H H B - B B A B H H H H B H A B B H H B B H B A H H H H H
B H A A H H B A B B H B H H H H B B H H B H B H H H H A H H B A H
A H H A B B B A A H H H H A B H H
```

```
individual names:
G00A
G00B
G003
G004
G005
G006
G007
G008
...
```

Figure 2.11: Middle section of the MapQTL file, `F2maize.loc`, showing specification of the genotype labels.

The map file (`F2maize.map`) contains two columns of data, split into sections corresponding to linkage groups. Each linkage group has a header of the form "group name", followed by the markers and their position within the linkage group. This format is illustrated in Figure 2.12.

```
group 1
L008       0
L058     7.4
L094    41.3
L040    73.4
L112    83.4
L062    84.8
L065   127.5
L085     141
L039   150.2
L082   174.8
L123   195.8
L070     206
L117   215.3
L032     227
L013   232.7
L049   241.6
L057     248
L028     252
L124     266

group 2
L023       0
L111    16.4
L120    55.4
L002      77
...
```

Figure 2.12: Top section of MapQTL file (`F2maize.map`) showing specification of the genetic map for the F2 population used in the CIMMYT maize trials.

2.1.2.4   Loading genotypic data

Genotypic data can be loaded from files either:

- via Stats | QTLs (Linkage/Association) | Import/Export | Load Genotypic (Marker and Map) data; or,
- by clicking on the Open QTL data file icon (📂) on the QTL Data View and selecting the option Genotypic (marker and map) data file.

Both routes open the Open Marker and Map Data files menu (Figure 2.13).

Figure 2.13: Loading genotypic data, for F2 population used in the CIMMYT maize trials, from Flapjack format files `F2maize_geno.txt` and `F2maize_map.txt`.

For Flapjack and R/qtl input files, it is necessary to specify the population type to one of the settings in Table 2-1. The population type is read from files in MapQTL® .loc format.

For Flapjack format files, it is possible to specify the delimiter used between data items, the separator used between parental alleles (offspring only) and the missing data string (see Options pane of Figure 2.13).

Table 2-1: Permitted population types for QTL mapping in Genstat 18th Edition.

Code	Description
F2	F2 offspring from F1 cross[†]
DH1	Doubled haploid offspring from an F1 cross[†]
BC1	Back-cross of parent to an F1 cross[†]
RILn	Recombinant inbred lines at generation n[†]
BCxSy	Back-cross inbred lines[†]
CP	Full-sibling family of outbreeders[‡]
Association mapping	Association mapping population

[†] two homozygous parents
[‡] two heterozygous parents

By default, all the data read will be loaded into the QTL Data Space. A summary of the imported data will be printed in the Output window, together with a report on any errors encountered. If errors occur, the default action is to abort the data import procedure so that incorrect data (which could lead to an incorrect analysis) is not loaded into the QTL Data Space. Details of the checks made are given in Section 2.1.2.5 below. You can then check and correct the marker and/or map data before re-loading. Alternatively, you can use the If markers contain errors: box to specify that the data should be imported after removing all markers found to contain errors. In this case, the error report should be carefully examined, paying particular attention to the list of markers removed.

A set of default names are suggested for the structures loaded into Genstat, however alternative names can be provided. The default names and an overview of their contents are listed in Table 2-2. The codes for the parental alleles are stored and then translated into a standard format. For example, for bi-parental populations arising from F1 crosses, the allele for the first parent is translated into code 1 and the allele for the second parent is translated into code 2. The offspring alleles are then also translated into this format, giving codes of the form 1/1, 1/2, 2/2 and so on (dependent on the population type). For each population type, this results in a reference set consisting of all possible offspring codes. The scores for each marker are then held as a factor, with the levels defined by the reference set, and these factors are all held in a single pointer. A set of genotype labels is created to identify the rows of the marker score factors with the individual genotypes; this structure is used to help verify the matching of genotypes across the genotypic and phenotypic data sets. The remaining structures hold information on markers (names,

linkage groups and positions within linkage groups) or parents (allele codes before translation and names of parents).

Table 2-2 Structures created on reading in genotypic data for a population with n individuals (excluding parental lines) and m markers

Structure	Default name	Type	Description
Marker genotypes	m_scores	Pointer	Pointer to m factors (one for each marker) each with n values. All factors have a common set of levels representing the reference set of parental allele combinations.
Marker names	m_names	Text	Text with m values, containing marker names.
Linkage groups	m_linkage	Factor	Factor with m values, identifying the linkage group for each marker.
Positions within linkage groups	m_positions	Variate	Variate with m values, containing map position within the linkage group for each marker.
Genotype labels	m_id	Text	Text with n values, containing genotype labels to be used for combining genotype and phenotype data sets.
Parental information	m_parent	Pointer	Pointer to population parents. Each element is a text with n values, defining the allele for each parent at each marker.
Parent labels	m_parentid	Text	Text with length equal to the number of parental lines, containing list of parent names.

Alternatively, marker score and map data in the required format (see Table 2-2) can be loaded directly into Genstat (e.g. using File | Open) and then transferred into the QTL Data Space by either: selecting the Add data in Genstat to QTL data space icon (⊞) on the QTL Data View, followed by the Genotypic data tab; or by right clicking on the *Genotypic data* folder in the QTL Data View, then selecting Add to data space.

If you have imported genotypic data from R/qtl or MapQTL formats, you can save the imported data in the Flapjack text file format via either Stats | QTLs (Linkage/Association) | Data Import/Export | Save Genotypic (Marker and Map) data or by clicking on the Save QTL data space icon (🖫) on the QTL Data View and selecting the option Genotypic (marker and map) data files.

2.1.2.5 Validation of genotype marker and map data

Various checks are carried out to ensure the basic integrity of the marker and map data. These checks are:

1. For inbred populations, that there are no errors in the parental alleles. For example, if neither a missing data code nor allele separator is used (i.e. only one allele supplied) an error results.
2. For mapping populations, the marker scores have the correct number of alleles.
3. For inbred populations, the marker scores have the correct number of alleles and only alleles specified in the parents or the missing data code are used.
4. That there are no duplicate marker names.

If any of these errors are found the marker and map data files will fail to load.

The data are also checked for a mismatch between the markers in the two files. If there are more markers in the genotype file than the map file, only those common between the two files will be loaded. All other markers are ignored. However, if the map file contains more markers than the genotype file, the marker and map data will fail to load.

2.1.3 Genetic relationship data

Genetic relationship data are used in association mapping analyses (Chapter 9) to indicate structure within the population. Within Genstat, this structure can either be estimated within the analysis (via eigenanalysis or by calculating a kinship matrix) or defined by an imported structure, either a kinship (or co-ancestry coefficient) matrix or a subpopulation grouping factor.

Genetic relationship data can be loaded from files via either:

- Stats | QTLs (Linkage/Association) | Data Import/Export | Load Genetic Relationship data; or,
- by clicking on the Open QTL data file icon (📁) on the QTL Data View and selecting the option Genetic relationship data files.

Both routes open the Open Genetic Relationship Data files menu (Figure 2.14), and by default the data loaded is added into the QTL Data Space.
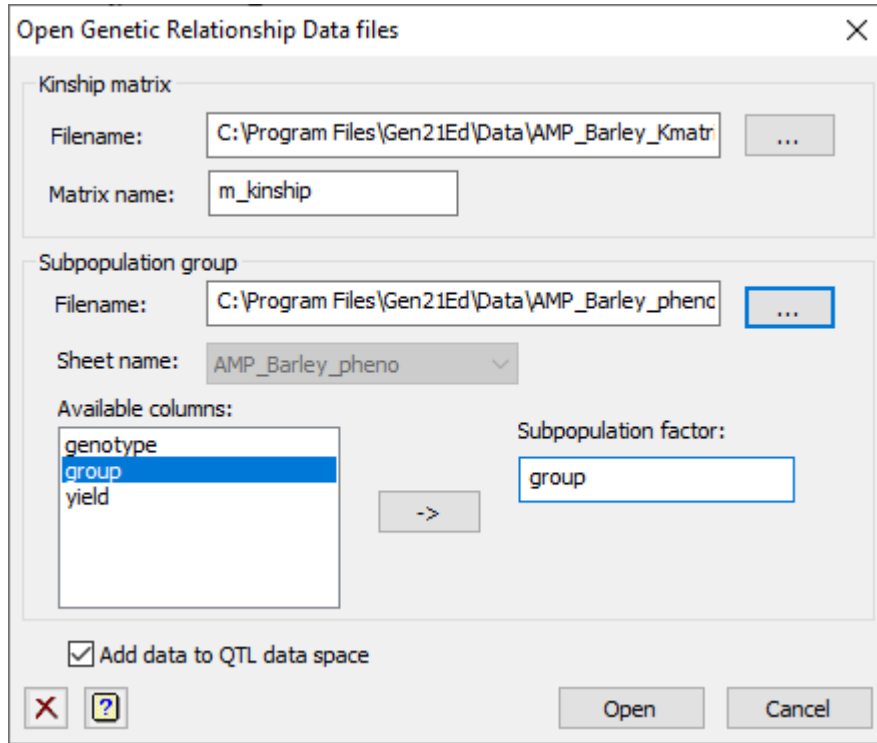
Figure 2.14: Importing genetic relationship data; kinship matrix (`AMP_Barley_Kmatrix.txt`) and subpopulation structure (`AMP_Barley_pheno.csv`).

The kinship matrix is a symmetric matrix with value 1 on the diagonal and co-ancestry coefficients ($0 \leq \theta \leq 1$) between all pairs of genotypes elsewhere. This type of matrix can be imported from a Genstat spreadsheet file (which must contain the structure as a symmetric matrix sheet) or from a text file. The first row of the text file contains the names of the genotypes, and the following rows the low-diagonal values of the coefficients of co-ancestry matrix between genotypes, diagonal included (for an example, see Figure 2.15). The file name must be specified in the Kinship matrix section of the Open Genetic Relationship Data files menu (Figure 2.14). For importing from a text file, the matrix name must be provided (default `m_kinship`); this name is read directly from a spreadsheet file. Figure 2.14 shows the import of a kinship matrix for the MABDE barley association panel (see Section 1.3.4) from text file `AMP_Barley_Kmatrix.txt`.

MABDE_001	MABDE_003	MABDE_004	MABDE_005	MABDE_007	MABDE_008	MABDE_009
1						
0.1793	1					
0	0	1				
0.115	0.1229	0	1			
0.2899	0.3126	0	0.121	1		
0.2946	0.1732	0	0.1869	0.3402	1	
0.2831	0.3517	0	0.1505	0.9018	0.3428	1
0.1561	0.389	0	0.0775	0.2268	0.2795	0.2443
0.2807	0.3271	0	0.1311	0.3373	0.3169	0.3387
0	0	0	0.1237	0.0313	0.0232	0.0068
0	0	0	0.1173	0.0325	0.028	0.0079

Figure 2.15: Top left portion of the kinship matrix for the MABDE barley association panel (Section 1.3.4) from text file `AMP_Barley_Kmatrix.txt`.

The Subpopulation factor: defines subsets of genotypes based on either some known population substructure or using the results of clustering the genotypes based on either genotypic or phenotypic data. For example, program STRUCTURE can be used to infer subpopulation structure and assign individuals to groups (see Pritchard *et al*., 2000, and website http://pritchardlab.stanford.edu/structure.html). The grouping can be imported from a Genstat spreadsheet, or Excel and text file formats for use in association mapping. The file name containing the grouping must be specified in the Subpopulation group section of the Open Genetic Relationship Data files menu (Figure 2.14), with the sheet name specified if multiple sheets are present in a spreadsheet (or Excel) file. Genstat spreadsheet, Excel files, and text files should contain the data structures in columns. The order of the genotypes in this grouping must match the order of the genotypes in the phenotypic data. Figure 2.14 imports a grouping factor (`group`) for the MABDE barley association panel (see Section 1.3.4) from the phenotypic text file `AMP_Barley_pheno.csv`.

Alternatively, genetic relationship data can be loaded directly into Genstat (e.g. using File | Open) and then transferred into the QTL Data Space by either: selecting the Add data in Genstat to QTL data space icon (⊞) on the QTL Data View, followed by the Genotypic data tab; or by right clicking on the *Genotypic data* folder in the QTL Data View, then selecting Add to data space. Selection of Association mapping in the Type of population: box activates the Subpopulation groups: and Kinship matrix: boxes that can be used to add these structures to the QTL Data Space.

2.2 QTL Data Space

Data structures imported into the QTL Data Space (introduced in Section 1.2) are then available in the QTL Data View (Figure 1.3). Symbols indicate the types of the data structures:

 factor

 variate

 text vector

 pointer

 symmetric matrix

For more information on different types of data structures, see the *Syntax and Data Management Guide* (Chapter 2). You can move the cursor over the structure names to obtain tooltip descriptions of their properties.

The QTL Data Space is a convenient way to manage data required for QTL analysis within Genstat. By default, Genstat will fill fields in the QTL menus using suitable structures from the QTL Data Space. You can also save a QTL Data Space and reload it later, ready to recommence analysis.

The QTL Data Space can be saved as a .qds file either: via Stats | QTLs (Linkage/Association) | Data Import/Export | Save QTL Data Space; or by clicking on the Save QTL data space icon ( ) in the QTL Data View; or by right clicking anywhere in the QTL Data View and selecting Save data space from the drop-down menu.

To (re-)load a QTL Data Space, you can use either: the Stats | QTLs (Linkage/Association) | Data Import/Export | Load QTL Data Space menu; or click on the Open QTL data file icon ( ) in the QTL Data View and select QTL data space; or right click anywhere in the QTL Data View and select Load data space from the drop-down menu.

To examine the values of structures in the QTL Data Space, right click on the group name (e.g. *PhenoMeans*) and select Create Spreadsheet from the drop-down menu. Genstat will open all structures in that group in spreadsheets, using separate sheets for structures of different lengths.

To remove all data structures from the QTL Data Space click on the Delete data from QTL data space icon ( ) in the QTL Data View. An information box (Figure 2.16) will appear advising you to choose either:

- Delete, which removes the data from the QTL Data Space and deletes it from the Genstat Server, or
- Remove, which removes the data from the QTL Data Space but retains it in the Genstat Server, or
- Cancel, which takes no action.

Alternatively to remove the data from the QTL Data Space (but retain it in the Genstat Server), right click in the QTL Data View and select Delete all from the drop down menu.

To remove a single data structure from the QTL Data Space, select this structure in the QTL Data View and use the Delete key on the keyboard. Figure 2.16 will appear, and you can select Delete or Remove as appropriate.
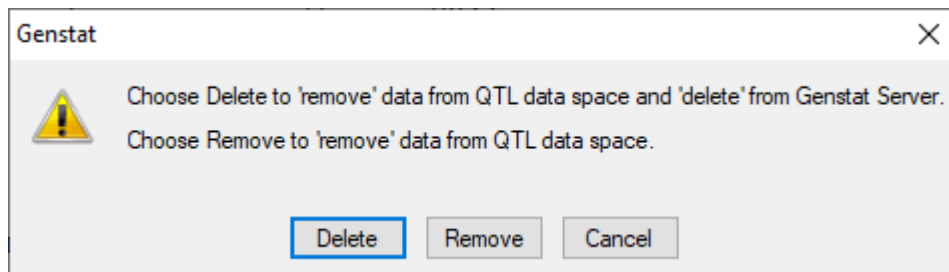
Figure 2.16: Options for deleting data from QTL Data Space.

It is also possible to save the contents of the QTL Data Space (data structures and results of analysis) into a Flapjack project file (Milne *et al*., 2010). You can do this using either: the Stats | QTLs (Linkage/Association) | Data Import/Export | Save Flapjack Project file menu; or click on the Save and open in Flapjack icon (⬛) in the QTL Data View. Both routes will open the Export to Flapjack project file menu, with options for creating a Flapjack project file displayed on different tabbed pages as follows:

- Flapjack project tab - options for saving the Flapjack project file.
- Genotypic data tab - options for specifying the genotypic (marker and map) data to be included in the Flapjack project file.
- Trait data tab - options for specifying the trait data to be included in the Flapjack project file.
- QTL data tab - options for specifying the results from a QTL analysis to be included in the Flapjack project file.
- Export files tab - options for specifying the files to form the Flapjack project file.

2.3 Data manipulation

In QTL experiments, it is quite common to either obtain phenotypic data for only a subset of a mapping population or to obtain marker information for only a subset of the individual genotypes. This can lead to a mismatch between the genotypes present in the phenotypic and genotypic data sets, and/or between the marker scores present for the genotypes and the markers present in the map. In addition, there may be a discrepancy between the ordering of the genotypes between the phenotypic means and genotypic data sets, which must be recognized and accommodated in order to achieve a valid analysis. In this section, we describe the checking and subsetting facilities available from the Data Manipulation menu that can be used to ensure the correct data sets (or subsets) are used for QTL detection. This menu can be accessed either via Stats | QTLs (Linkage/Association) | Data Manipulation; or, from the Manipulate button on the QTL Data View. We first describe the Compatibility Check menu that is used to check consistency across the phenotypic and genotypic data sets (Section 2.3.1), and exclude genotypes or markers with poor quality. Once a coherent data set has been established, then further subsetting may be desirable, and menus for subsetting are described in Section 2.3.2.

2.3.1 Compatibility across phenotypic and genotypic data sets

The Compatibility Check menu is designed to be used before QTL analysis and performs a range of checks required to obtain a coherent combined data set. Variables from the genotypic and phenotypic data sets are used as input, and the contents of the QTL Data Space will be used to initialize the menu. For each input variable, a new name must be provided for the corresponding (possibly) reduced output variable. By default, these output variables will replace the original variables in the QTL Data Space; this can be changed using the check box on the Compatibility Check Options sub-menu (opened by clicking on the Options button).

Within the genotypic data set, the procedure will check for, and exclude, any markers with <50% scores present or with <5% frequency of any one allele. These thresholds can be changed using the Compatibility Check Options settings Percentage of missing values allowed in markers: and Extreme allele percentage allowed for marker (between 0 and 5):, respectively. Setting these options to zero will mean that no markers are excluded on the basis of missing values or allele frequency.

Labels of the genotype factor (as part of the phenotypic data set) will be checked against the set of *Genotype labels* (part of the genotypic data set). The procedure will exclude any genotypes not present in both data sets. In addition, genotypes with scores

present for <50% of the marker set will be excluded. This threshold can also be changed using the Compatibility Check Options setting Percentage of missing values allowed for genotype:. Setting this option to zero will mean that no genotypes are excluded on the basis of missing marker scores. Finally, the new genotypic data variables will be reordered so that the *Genotype labels* match the ordering of the labels of the *Genotype factor*, to ensure that the two data sets can be combined without error.

2.3.2 Subsetting

Additional subsetting of phenotypic or genotypic variables may also be required, for example, to restrict QTL analysis to a single chromosome. Within the QTL Data Space, subsetting can be reversed and/or revised, if necessary, as backups of the full data sets are stored. The full data can be restored by selecting the Stats | QTLs (Linkage/Association) | Data Manipulation | Remove Subsetting menu item, or by using the Manipulate button on the QTL Data View.

2.3.2.1 Subsetting phenotypic and genotypic data by genotypes

Subsetting by genotypes can be used to remove parents of a population from the trait means and genotypic data or to remove genotypes with unreliable marker scores or trait means. Subsetting by genotypes should only be done once compatibility of the phenotypic and genotypic data sets has been established (Section 2.3.1). The Subset Phenotypic and Genotypic Data by Genotypes window is accessed from the Data Manipulation menu and gives a list of all genotypes in the data set (Figure 2.17).

All genotypes to be retained in the subset should be copied to the right-hand pane. The Select all button allows the full set of genotype names to be copied across; this can be useful if only a few genotypes are to be excluded. Checking the box Delete genetic predictors and associated information will delete any genetic predictors in the QTL Data Space (i.e. genetic covariates associated with marker information, see Section 6.1.1). These can be recalculated prior to QTL analysis using only the genotypes retained in the subset (see Chapter 6).

After subsetting by genotypes, the trait means and genotypic data in the QTL Data Space will only contain units corresponding to the selected genotypes. The number of genotypes in the subset is shown by placing the mouse cursor over the genotype factor name (default `genotype`) in the *PhenoMeans* section (as Nlevels) or over the genotype labels (default `m_id`) in the *Genotypic* section (as Nvals) of the QTL Data View.
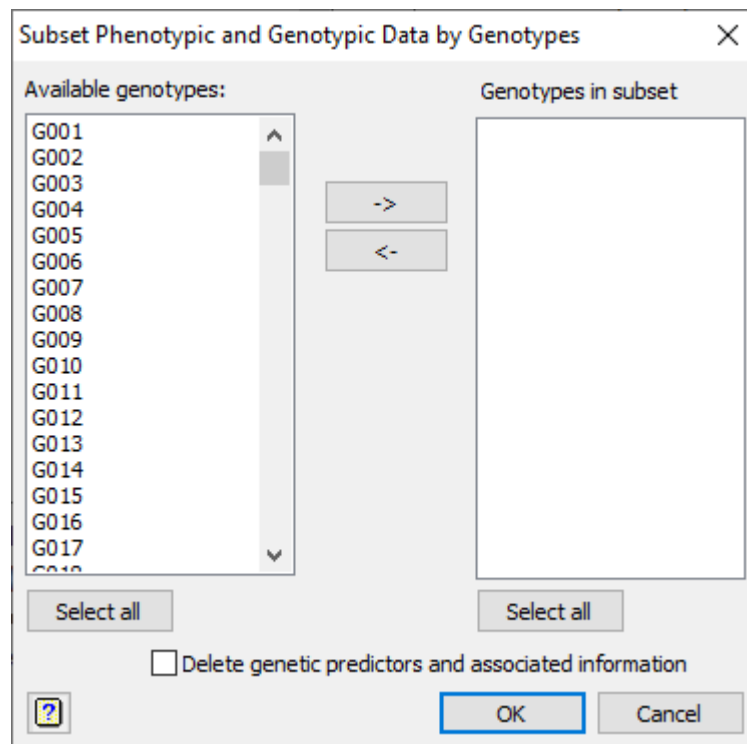
Figure 2.17: Subset Phenotypic and Genotypic Data by Genotypes window.

2.3.2.2 Subsetting the genotypic data by markers

Subsetting by markers can be used investigate a subset of the genome (e.g. one linkage group) or to exclude markers of poor quality. The Subset Genotypic Data by Markers window is accessed from the Data Manipulation menu and gives a list of all linkage groups and markers in the genotypic data set (Figure 2.18).

The set of markers to be included can be made by linkage group and/or by individual markers. Moving linkage groups into/out of the selected subset will cause all of the markers in that linkage group to be moved accordingly. Checking the box Delete genetic predictors and associated information will delete any genetic predictors that have been stored within the QTL Data Space. These can be recalculated prior to QTL analysis using only the markers retained in the subset (see Chapter 6).

The genotypic data variables in the QTL Data Space (linkage groups, marker scores, marker names and marker positions) will then only contain units corresponding to the selected markers. The number of linkage groups (Nlevels) and markers (Nvals) in the subset will be shown if you place the mouse cursor over the linkage group factor name (default `m_linkage`) found in the *Genotypic* section in the QTL Data View.
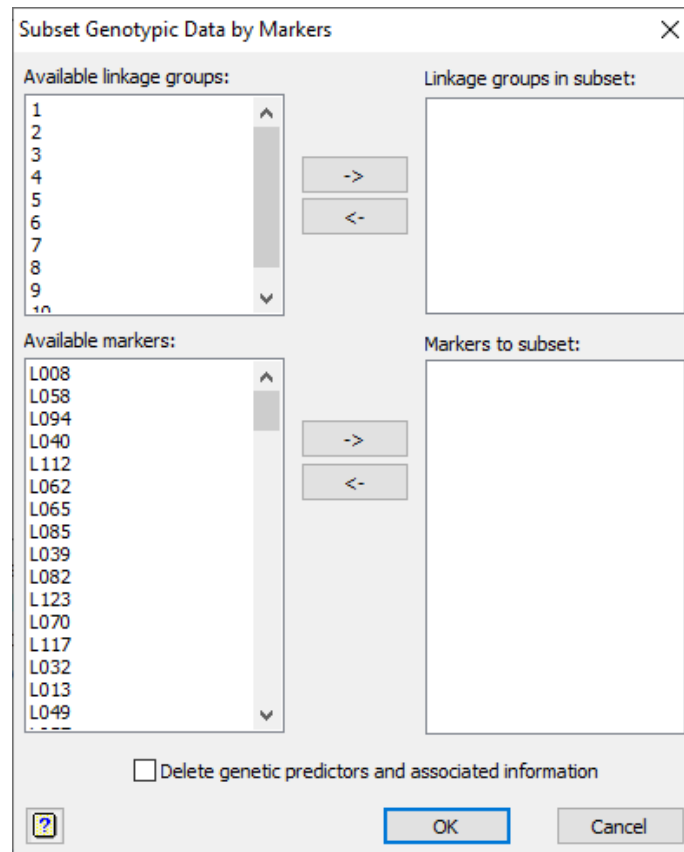
Figure 2.18: Subset Genotypic Data by Markers window.

2.4 Data exploration

Checking and exploration of data before analysis allows unusual observations, such as outliers or typographical errors, to be identified and (where possible) corrected. The QTL menu has a number of exploratory tools (summary statistics and graphical displays) that enable you to explore both phenotypic and genotypic data before embarking on QTL analysis. You can navigate to the data exploration tools via the menu Stats | QTLs (Linkage/Association) | Data Exploration or by using the Explore button in the QTL Data View.

2.4.1 Exploration of phenotypic data

Exploration of the phenotypic data sets is enabled by selecting the Phenotypic Data option under the Data Exploration menu. Here we will only describe the tools for producing summary statistics; the AMMI and GGE tools for exploration of multi-environment trials are discussed in Chapter 4.

2.4.1.1 Summary statistics by environment

The Summary Statistics menu allows you to calculate summary statistics for the selected trait within each environment via the Stats | QTLs (Linkage/Association) | Data Exploration | Phenotypic Data | Summary Statistics Single Environment menu. The summary statistics can be calculated on either the raw plot (unit) data or the trait means, depending on the trait variates selected. To calculate the summary statistics for each environment, specify the environment factor that corresponds to your trait variate (i.e. relating to either raw data or trait means for each genotype) in the By groups: box. The Display settings are used to select the summary statistics calculated for each environment, and the Graphics settings can be used to display the distribution of the data: the settings Histogram and Boxplot may be helpful in this context. For the CIMMYT maize trials (Section 1.3.2), Figure 2.19 shows histograms of yield (trait mean `yld`) from each environment: there is a separate histogram for each environment, labelled at the top of each plot. This shows a clear change in both the mean and variation of yields across the eight environments.
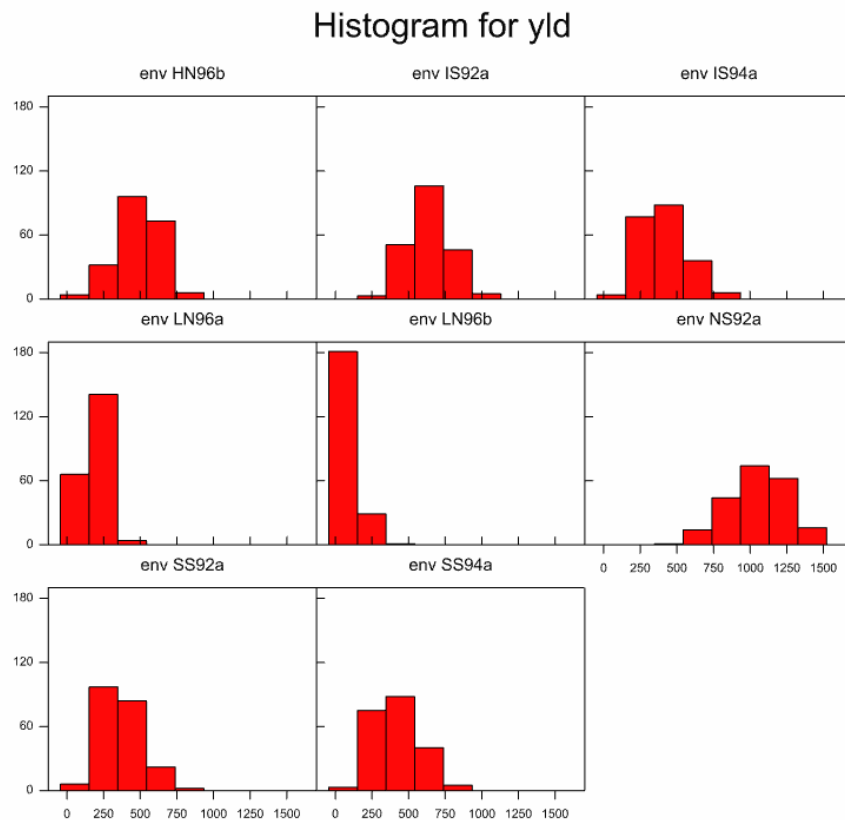
Figure 2.19: Histograms showing distribution of trait mean `yld` within each environment of the CIMMYT maize trials data set (Section 1.3.2).

Figure 2.20 shows boxplots of the same data set, with the distributions for each environment laid out side by side. The differences in mean and variance between environments are still clear, but the boxplots also identify a number of outlying values within each environment that were not visible in the histograms. These observations should be checked for errors before analysis.
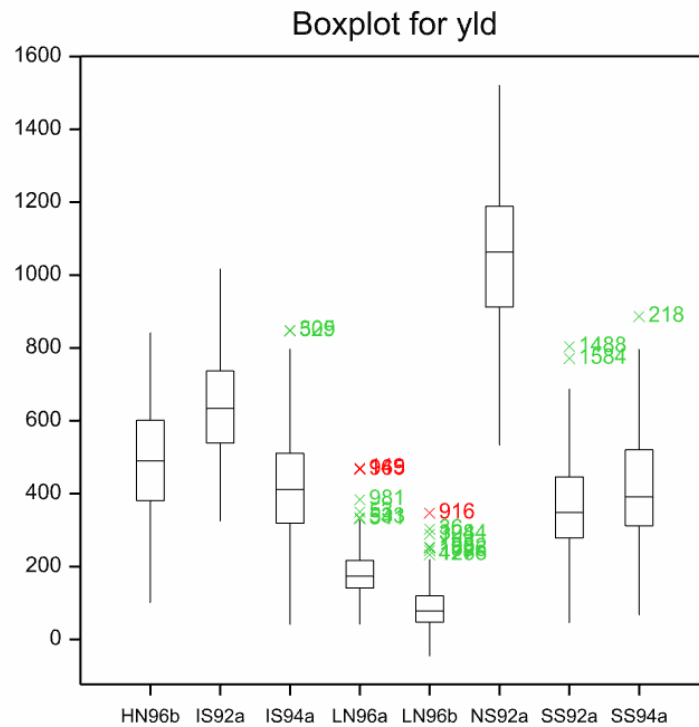
Figure 2.20: Boxplot showing distribution of `yld` within each environment of the CIMMYT maize trials data set.

2.4.1.2 Summary statistics between environments

The Summary Statistics Between Environments menu allows you to explore correlations between measurements on a single trait made in different environments. It is accessed via Stats | QTLs (Linkage/Association) | Data Exploration | Phenotypic Data | Summary Statistics Multiple Environments. Calculations can be done using either raw data or trait means. By default, Genstat will fill the menu using structures from the *PhenoMeans* section of the QTL Data Space, but these can be replaced by structures from the *Phenotypic raw data* folder if required. The form of the menu is shown in Figure 2.21.
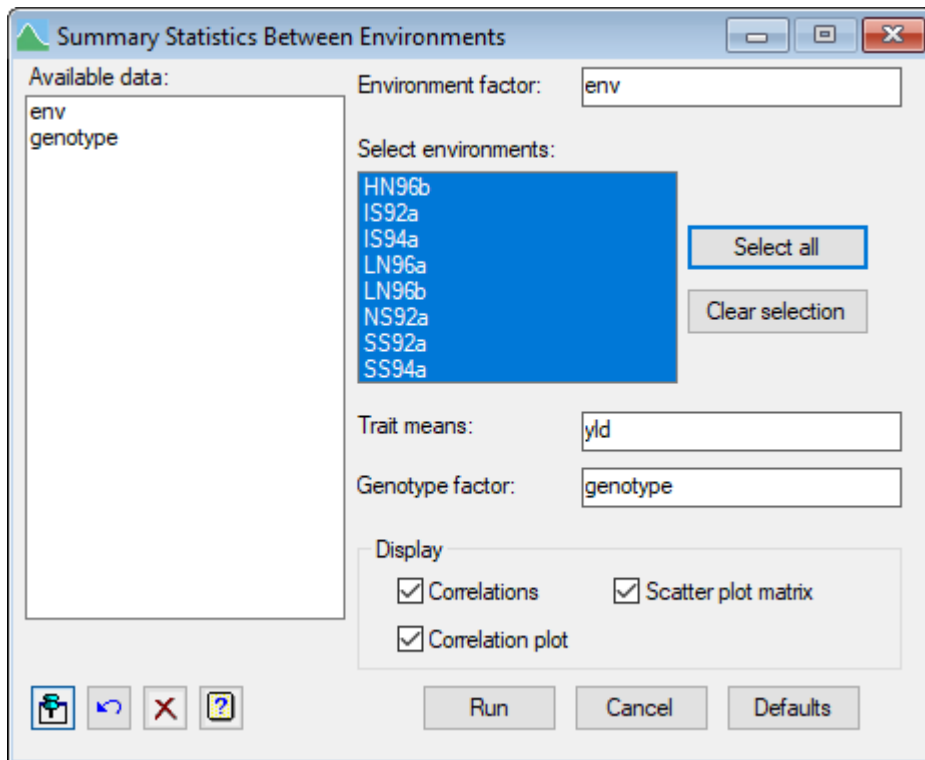
Figure 2.21: Summary Statistics Between Environments menu.

It is necessary to select the environments for which correlations are to be calculated or plotted, then choose which output to display. If the option Correlations is chosen, then an across-environment correlation matrix is printed in the Output window. If the option Correlation plot is chosen, then the correlation matrix values are displayed graphically, as shown in Figure 2.22, as a shade plot of the correlation matrix, coloured by spectrum from blue (correlation = -1.0) to red (correlation = +1.0). The environments are sorted alphabetically, coded by integers and these codes are used as the tick labels on the plot. The legend in the top right corner of the plot shows the correspondence between code and environment name. The option Scatter plot matrix produces a scatter plot for each pair of environments. This is best restricted to a small number of environments, otherwise the individual plots become too small to be useful. Figure 2.23 shows a scatter plot matrix for the five environments used in 1992 and 1994 for the CIMMYT maize trials. These plots provide a useful way to identify genotypes that behave differently between environments.
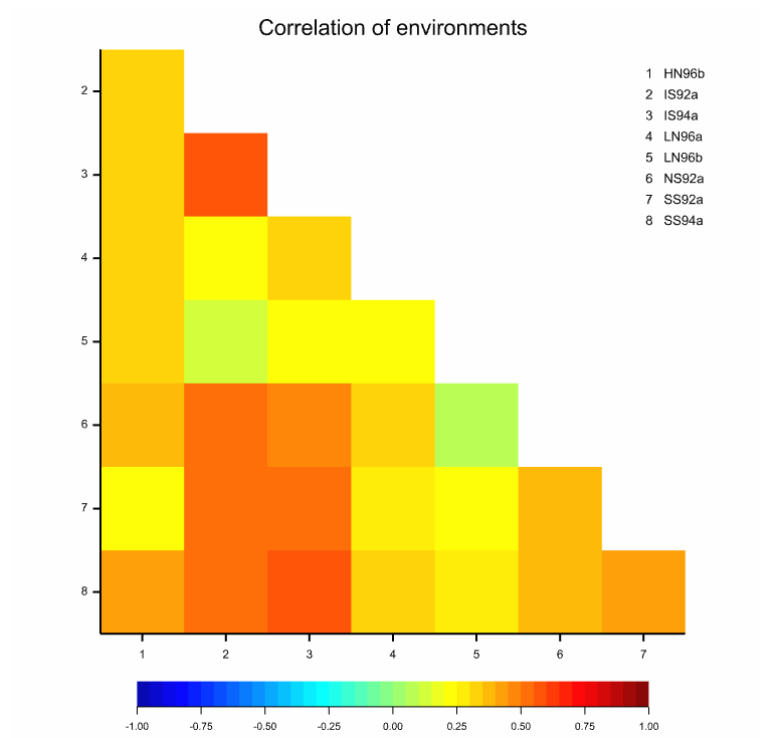
Figure 2.22: Correlation plot of `yld` across the 8 CIMMYT maize trial environments.
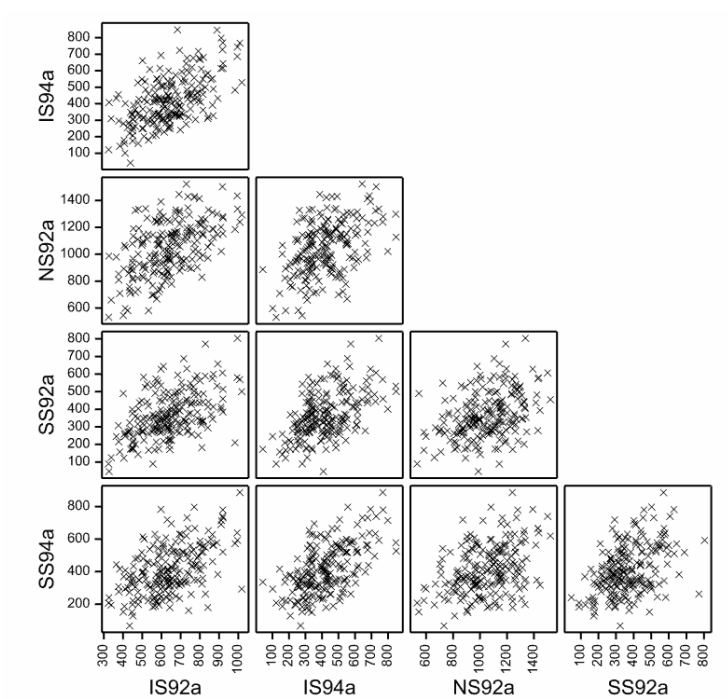
Figure 2.23: Scatter plot matrix of `yld` from the CIMMYT maize trials in 1992 and 1994.

2.4.1.3   Summary statistics between traits

The Summary Statistics Between Traits menu allows you to explore correlations between measurements on several traits made in a single environment, or measurements on two traits average over multiple environments. It is accessed via Stats | QTLs (Linkage/Association) | Data Exploration | Phenotypic Data | Summary Statistics Multi-traits. Calculations can be done using either raw data or trait means. By default, Genstat will fill the menu using structures from the *PhenoMeans* section of the QTL Data Space, but these can be replaced by structures from the *Phenotypic raw data* section if required.

It is necessary to select the traits for which correlations are to be calculated or plotted, and the environment for which calculations are to be formed, under Calculate statistics for:. The default setting of average will average the traits over all environments and then derive correlations from these averaged values. You can then choose which output to display. If the option Correlations is chosen, then an across-trait correlation matrix is printed in the Output window. If the option Correlation plot is chosen, then the correlation matrix is displayed graphically, similarly to Figure 2.22, but in this case the traits are sorted alphabetically. The option Scatter plot matrix produces a scatter plot for each pair of traits, similarly to Figure 2.23. The option Biplot produces a plot of the first two principal components calculated from the across-trait correlation matrix; this is discussed further in Chapter 4.

To explore the correlations between traits measured within a single environment, specify the environment of interest in the Calculate statistics for: field.

2.4.2   Exploration of genotypic data

Exploration of the genotypic data sets is enabled by selecting the Genotypic Data option under the Data Exploration menu. The tools for exploration data enable you to investigate the quality of the genotypic data, including understanding the genome coverage, the proportion and patterns of missing marker data, and the segregation ratios.

2.4.2.1   Display a genetic map

The Display Genetic Map menu produces a simple line plot of the genetic map that allows you to view the molecular marker density and coverage across the genome. If structures containing the allocation to linkage groups (or chromosomes), marker positions and marker names have been loaded into the QTL Data Space, then the menu will be populated automatically (Figure 2.24).
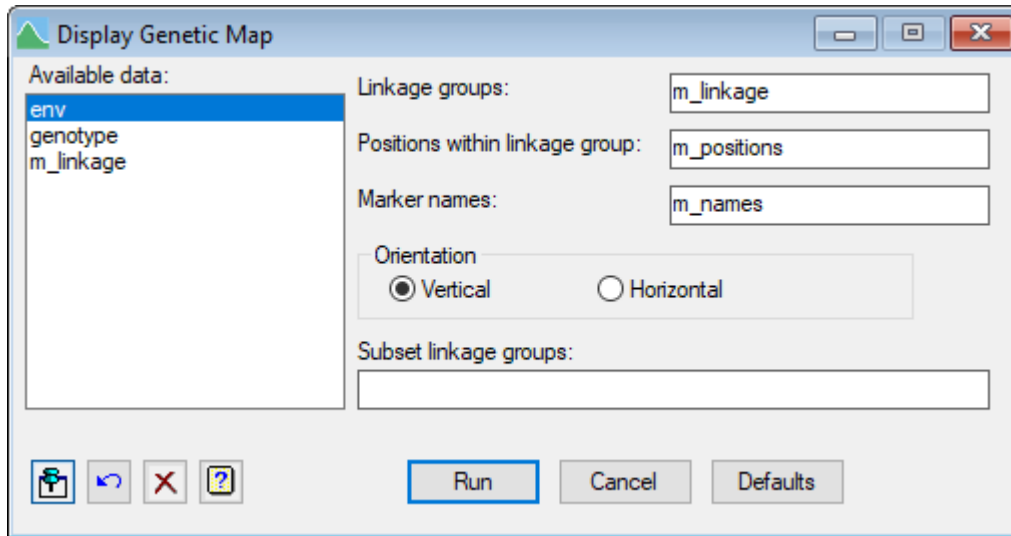
Figure 2.24: Display Genetic Map menu.

The Orientation used for the linkage groups (or chromosomes) can be Vertical (default, with linkage groups running down the page) or Horizontal (linkage groups running across the page). You can Subset linkage groups (or chromosomes) by supplying a comma separated list of the group number(s) or labels, e.g. 1, 2, 3 or A1, A2, A3, or by providing a variate or text structure containing the required group numbers or labels. The plot is displayed in the Genstat Graphics Viewer and the markers are labelled by tooltips using their names. To identify a marker select the Data Info tool (⟦🔎⟧) and then click on the marker of interest (see Figure 2.25).
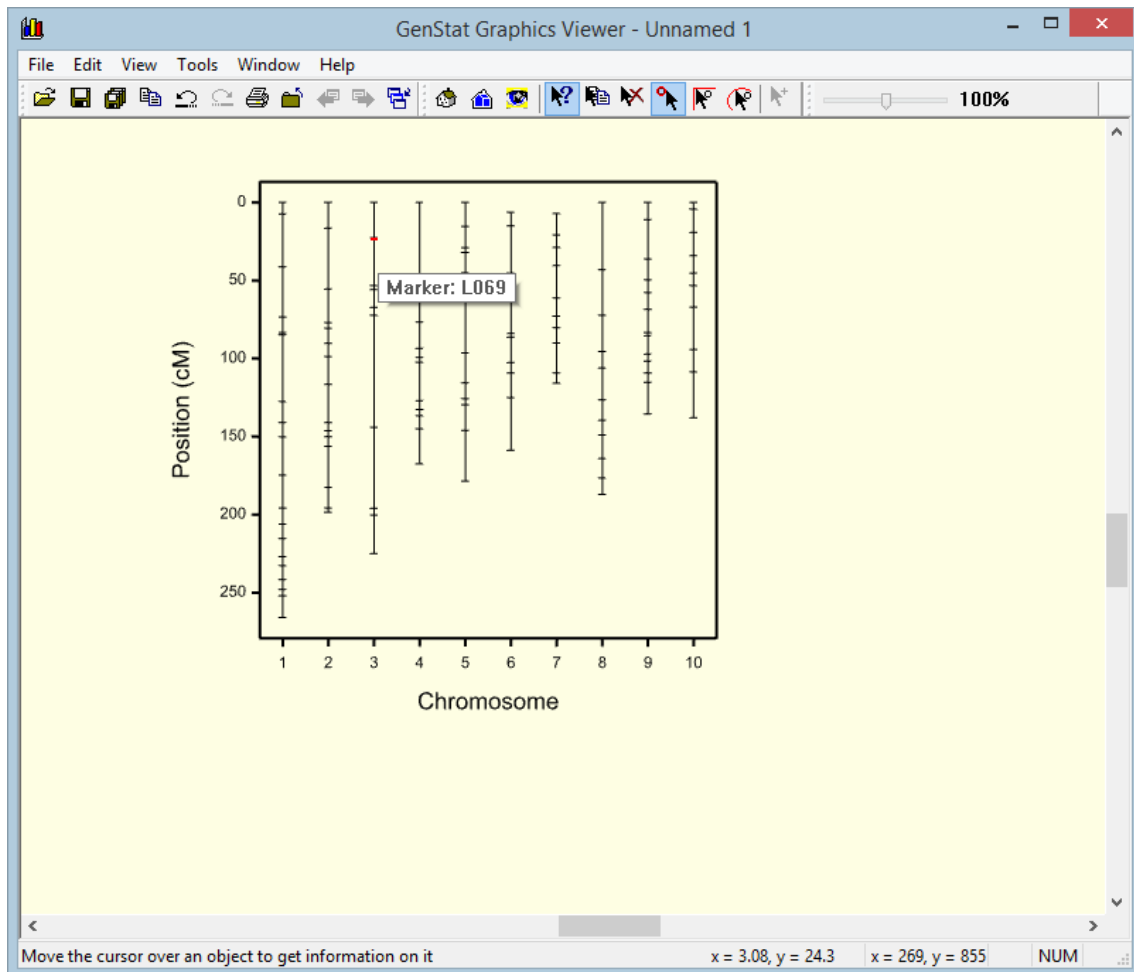
Figure 2.25: Using the Data Info tool to identify a marker in a genetic map.

2.4.2.2   Genotype data plots

Genotype data plots are simple graphical displays that assist you in visualising the marker data. The Genotype Data Plots menu automatically fills with structures from the QTL Data Space, if these are available (Figure 2.26).

Selecting All genotype scores (default) produces a shade plot of the genotype (rows) by marker (columns) matrix of marker scores, using different colours for each combination of parental alleles. Selecting Missing genotype scores produces a shade plot that highlights the missing marker scores, using different grey shades to differentiate between fully and partially missing marker information. In both plots, the default colour scheme can be changed via the Colours button. A row (i.e. genotype) subset of the matrix can be obtained by specifying a range of genotype numbers, using the Lower genotype: and Upper genotype:

boxes. A column (i.e. marker) subset can be specified by using the Subset linkage groups to supply a comma separated list of the group number(s) or labels, e.g. 1, 2, 3 or A1, A2, A3, or by providing a variate or text structure containing the required group numbers or labels.
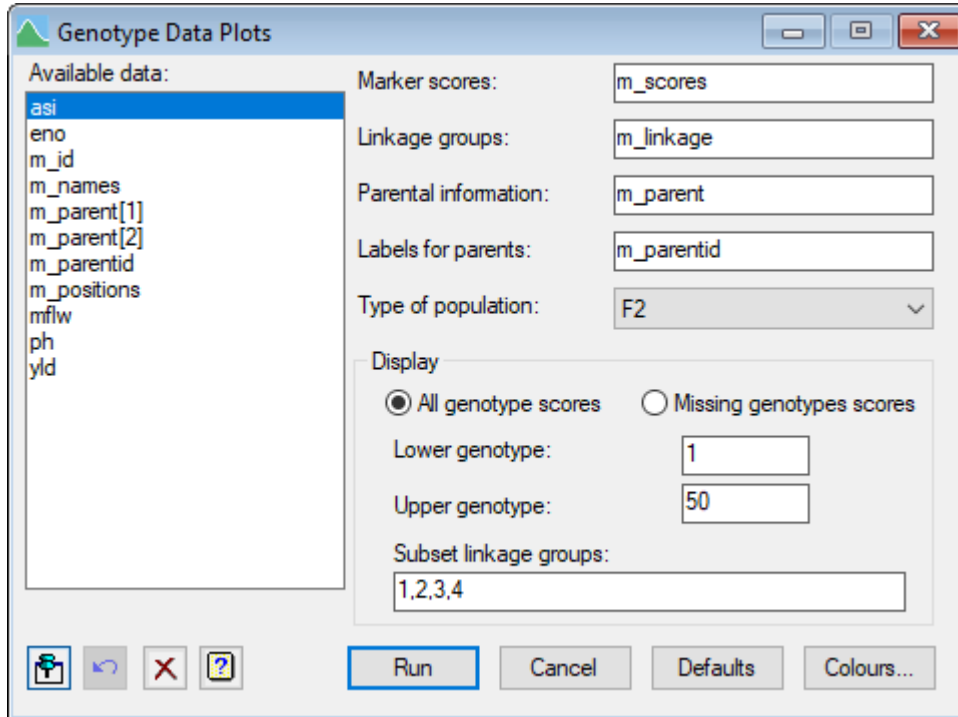
Figure 2.26: Genotype Data Plots menu. Here, for the CIMMYT maize trials, a plot for genotypes 1-50 at linkage groups 1-4 is requested.

Figure 2.27 gives the genotype data plot for genotypes 1-50 of the F2 CIMMYT maize population on linkage groups 1-4. In general, within linkage groups, we expect to see runs of reasonable length of the same colour, with occasional changes. There are clearly a few markers in linkage groups 1-4 with a large proportion of missing values (indicated by black squares).
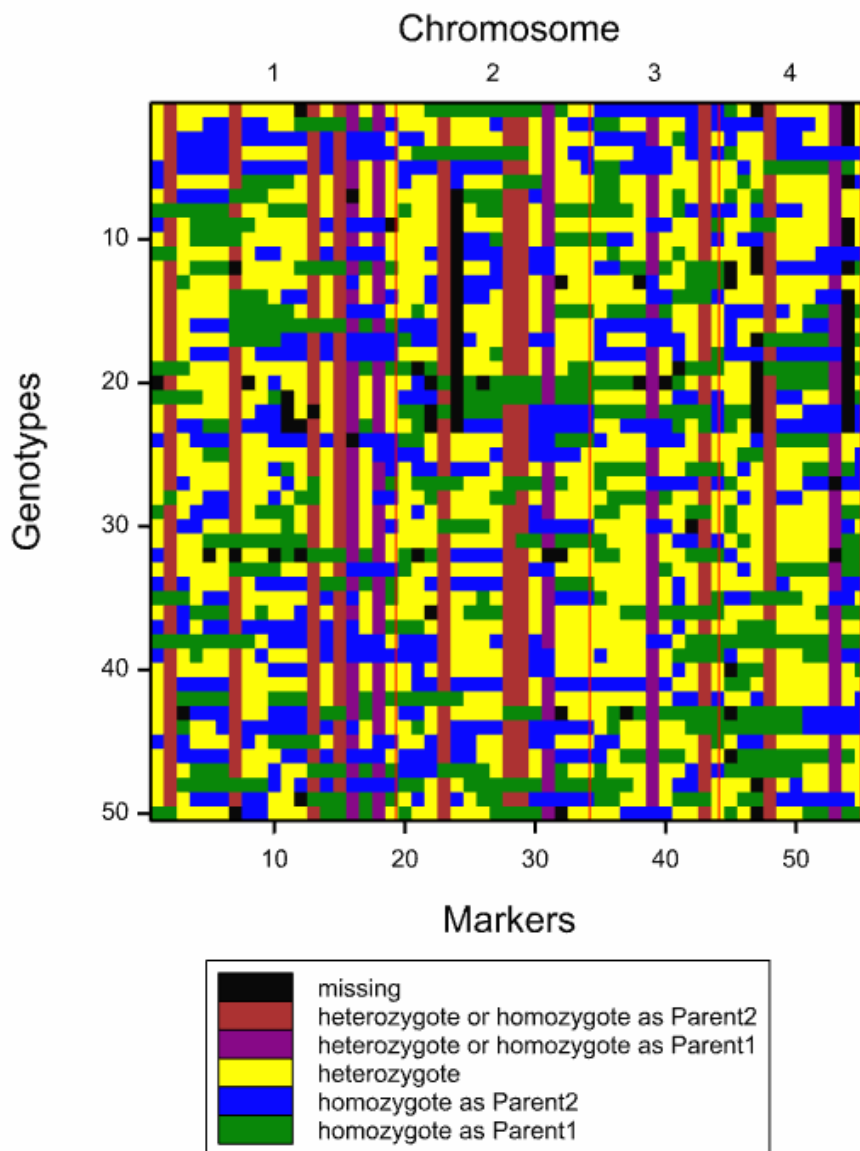
Figure 2.27: Genotype data plot, across linkage groups 1-4, for genotypes 1-50 of the CIMMYT maize trials.

2.4.2.3 Summary statistics for markers

The Summary Statistics for Markers menu produces summary information on the marker scores and genetic map. The data fields in the menu are automatically filled using structures from the QTL Data Space, if these are available (Figure 2.28).
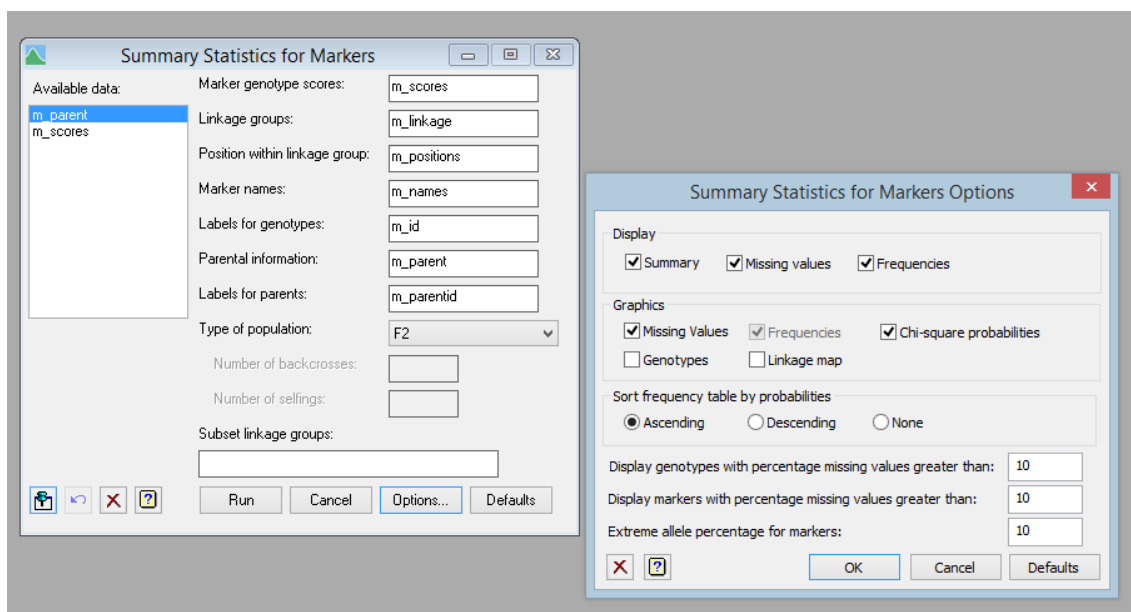
Figure 2.28: Summary Statistics for Markers menu with Options window. Here, summary statistics for the CIMMYT maize trials marker data is requested.

The display options, controlled by the Display section of the Summary Statistics for Markers Options menu (Figure 2.28), can be used to give a summary of the genetic map (setting Summary), a report on the pattern of missing values (setting Missing values) and an assessment of allele frequencies (setting Frequencies).

The genetic map summary reports the population type, the numbers of markers and genotypes, and parental names. Also reported, for each linkage group and the whole genome, are the number of markers, the total length and the median and 95th percentile of the inter-marker distances.

```
Summary
-------


            Population: F2
  Number of genotypes: 211
    Number of markers: 122

The labels of the parents are:
    Parent1
    Parent2


Chromosome     Length   Number of   Median distance   95% percentile
                        markers   between markers     of distances
```

```
         1        266.0         19          10.1           39.2
         2        198.5         15          11.4           36.5
         3        225.0         10          22.5           72.8
         4        167.5         11          12.6           48.3
         5        178.5         12          15.3           33.8
         6        152.5         11          13.2           33.7
         7        108.8         10          11.6           20.9
         8        187.0         11          14.1           43.1
         9        135.6         13          10.9           24.6
        10        138.0         10          14.3           29.5

    Genome       1757.4        122          12.8           38.5
```

A simple line plot of the genetic map (as described in Section 2.4.2.1) can be obtained by checking the Linkage map box under Graphics on the Summary Statistics for Markers Options menu (Figure 2.28).

The report on missing marker scores displays the total number of missing scores, the number of markers and genotypes with missing scores, and details of the markers and genotypes with >10% missing scores. The default threshold of 10% can be changed in either case through the Summary Statistics for Markers Options menu (Figure 2.28). A scatter plot of the number of missing scores against genome position and a shade plot of the missing value pattern (as described in Section 2.4.2.2) can be obtained by checking the Missing Values setting under Graphics on the Summary Statistics for Markers Options menu.

```
Missing values
--------------


There are 603 scores missing. This is 2.342% of the 25742 scores.


There are 86 markers with missing values. This is 70.49% of the 122 markers.

The 4 markers with more than 10% missing values over the 211 genotypes are:

                                  Number of       Percentage
Marker Chromosome    Position  missing values  missing values
  L055           4        76.7             27            12.8
  L080           4       145.0             30            14.2
  L115           6        45.2             35            16.6
  L127           8       106.0             32            15.2


There are 157 genotypes with missing values. This is 74.41% of the 211
genotypes.
```

```
The 8 genotypes with more than 10% missing values over the 122 markers are:

            Number of       Percentage
Genotype  missing values  missing values
   G020            20           16.4
   G023            14           11.5
   G032            20           16.4
   G063            21           17.2
   G075            17           13.9
   G121            13           10.7
   G128            14           11.5
   G133            16           13.1
```

The behaviour of the Frequencies option depends on the population type. For bi-parental mapping populations, a summary of the observed allele frequency (segregation ratio) is printed out for each marker, and a chi-square test is used to compare the observed frequencies with those expected for the population type. This test can be used to detect both systematic segregation distortion and suspect markers. The observed probability, P, (Prob.) under the null hypothesis (segregation as expected) is printed, and the markers are ordered by this probability value, with the lowest (most significant) first. This ordering can be changed using the Sort frequency table by probabilities section of the Summary Statistics for Markers Options menu.

```
Frequencies of markers (sorted on probabilities) with probability < 0.1
-----------------------------------------------------------------------


Marker  Chromosome    1/1    1/2    2/2    1/-    2/-    -/-    Prob.
  L031           9     45     94     71      0      0      1    0.013
  L115           6     35     82     59      0      0     35    0.025
  L007           4     45    123     41      0      0      2    0.035
  L035           3     58    114     36      0      0      3    0.037
  L094           1     36    116     55      0      0      4    0.039
  L087           8      0      0     38    165      0      8    0.039
  L098           6     40      0      0      0    171      0    0.043
  L016          10     37    116     58      0      0      0    0.043
  L020           3     59    109     36      0      0      7    0.046
  L100           3      0      0     40    170      0      1    0.046
  L078           7     68     98     45      0      0      0    0.048
  L108           5     40      0      0      0    169      2    0.050
  L061          10     37    116     54      0      0      4    0.055
  L125           5      0      0     64    147      0      0    0.074
  L104           6     41    106     64      0      0      0    0.081
  L024           9     36    103     57      0      0     15    0.082
  L070           1     39    117     50      0      0      5    0.083
```

A scatter plot of -$\log_{10}$(P) against genome position (Figure 2.29) can be obtained by checking the Chi-square probabilities box under the Graphics section of the Summary Statistics for Markers Options menu. Large values of -$\log_{10}$(P) (where 1.3 is equivalent to P=0.05 and 2.0 is equivalent to P=0.01) may indicate the presence of segregation distortion.
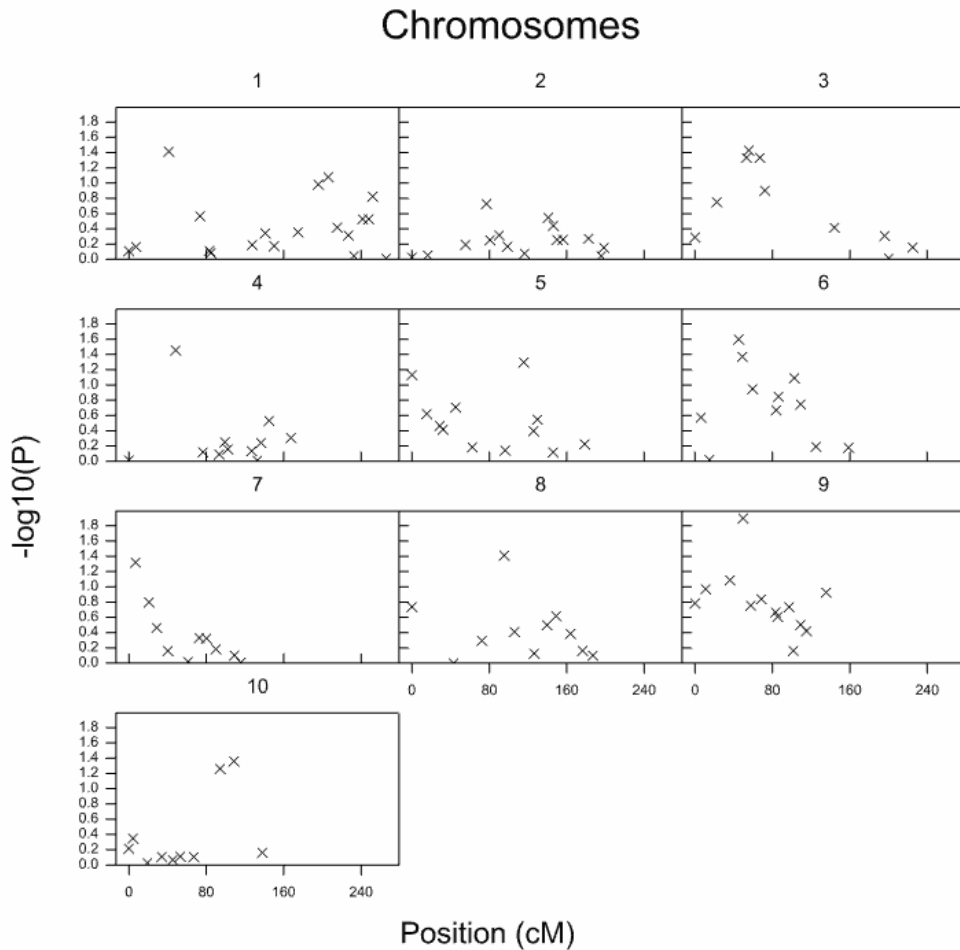
Figure 2.29: Plot of -$\log_{10}$(P) from the chi-square test for segregation distortion against genome position for the F2 maize population used in the CIMMYT trials.

For association mapping populations, a summary of the markers with rare alleles is also printed, where a rare allele is present for <10% of the genotypes. This threshold can be changed using the Extreme allele percentage for markers: box on the Summary Statistics

for Markers Options menu. The frequency of the rarest allele at each marker can be plotted against genome position by checking the Frequencies box under the Graphics section of the Summary Statistics for Markers Options menu.

All of this information can be obtained for a subset of the linkage groups (or chromosomes) by supplying a comma separated list of the group number(s) or labels, e.g. 1, 2, 3 or A1, A2, A3, or by providing a variate or text structure containing the required group numbers or labels in the Subset linkage groups: box of the Summary Statistics for Markers menu (Figure 2.28).

2.5   References

Broman, K.W., & Sen, Ś. (2009). A guide to QTL mapping with R/qtl. Springer, New York.

Milne, I., Shaw, P., Stephen, G., Bayer, M., Cardle, L., Thomas, W.T.B., Flavell, A. J., & Marshall, D. (2010). Flapjack - graphical genotype visualization. *Bioinformatics*, **26**, 3133-3134.

Pritchard, J.K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945-959.

van Ooijen, J.W. (2009). MapQTL® 6, Software for the mapping of quantitative trait loci in experimental populations of diploid species. Kyazma B.V., Wageningen, Netherlands, http://www.kyazma.nl/index.php/mc.MapQTL.

3 Preliminary phenotypic analysis: producing trait means per genotype from trial data

QTL analysis in Genstat requires a single value for each genotype (or line) in each environment for each trait (known as "*trait means*"), rather than the raw data values from replicated trials. It may therefore be necessary to perform an appropriate preliminary analysis to acquire trait means. This will usually involve analysing the full experiment according to the experimental design, and saving the predicted means for those genotypes to be included in the QTL analysis. The Preliminary Single Environment Analysis (PSEA) menu helps you to produce trait means from experimental data. This menu is not designed for a full analysis of an experiment, which should include model selection and storage of results. This should be done beforehand, for example by using the appropriate ANOVA or mixed model (REML). The PSEA menu takes a model that you have already established, and uses it to obtain suitable predicted trait means for QTL analysis.

In this chapter, we first illustrate the use of the PSEA menu and describe the underlying methods, then take a step back and provide some guidance on the process of model selection that should precede use of this menu. We first perform a standard single trial analysis (Section 3.1) to produce and save trait means and unit errors (i.e. weights) for use in subsequent G×E and QTL analyses (Section 3.2). We describe how to calculate the heritability of a trial (Section 3.3), and how to deal with genotypes that will not be used in the QTL analysis (Section 3.4). We compare different designs enabled on the PSEA menu (Section 3.5) and, finally, outline the process of model selection (Section 3.6).

In this chapter you will learn:

- how to use the Preliminary Single Environment Analysis menu (Section 3.1)
- how to generate and save the table of means and unit errors for subsequent G×E and QTL analyses (Section 3.2)
- how to calculate the heritability for the trial (Section 3.3)
- how to structure the analysis when not all genotypes are to be used in QTL analysis, e.g. when control or parent lines are present (Section 3.4)
- how to select models working from simple experimental designs to more complex modelling of spatial variability (Sections 3.5 and 3.6)
- how to produce and interpret variograms (Section 3.6)

3.1   Preliminary single environment analysis

We will first illustrate the process of obtaining trait means and unit errors using the raw plot yields from the CIMMYT spring wheat trials data set (introduced in Section 1.3.3), before discussing some more detailed features of the menu in the following sections. This data set contains raw plot yield data for a recombinant inbred line population, grown in alpha-lattice designs with two replicates in four environments (factor `Env`) by the International Centre for Maize and Wheat Improvement (CIMMYT). These trials tested 169 lines of spring wheat (factor `Genotype`). The design consists of two replicates (factor `Rep`), each containing 13 sub-blocks (factor `Subblock`) with 13 plots. The field layout for trial `HEAT05` is shown in Figure 3.1. The first 13 rows are the first replicate, and rows 14-26 are the second replicate. Each row within each replicate is a sub-block. Import the phenotypic data, held in file `SB_yield.csv`, as described in Section 2.1.1.1.

1	2	3	4	5	6	7	8	9	10	11	12	13
1 SB001	2 SB002	3 SB003	4 SB004	5 SB005	6 SB006	7 SB007	8 SB008	9 SB009	10 SB010	11 SB011	12 SB012	13 SB013
26 SB028	25 SB027	24 SB026	23 SB025	22 SB024	21 SB023	20 SB022	19 SB021	18 SB020	17 SB019	16 SB018	15 SB016	14 SB014
27 SB029	28 SB030	29 SB031	30 SB032	31 SB036	32 SB038	33 SB039	34 SB040	35 SB041	36 SB043	37 SB044	38 SB045	39 SB046
52 SB063	51 SB062	50 SB061	49 SB058	48 SB057	47 SB055	46 SB054	45 SB053	44 SB052	43 SB050	42 SB048	41 SB048	40 SB047
53 SB064	54 SB065	55 SB066	56 SB067	57 SB068	58 SB070	59 SB071	60 SB072	61 SB073	62 SB074	63 SB076	64 SB077	65 SB078
78 SB091	77 SB090	76 SB089	75 SB088	74 SB087	73 SB086	72 SB085	71 SB084	70 SB083	69 SB082	68 SB081	67 SB080	66 SB079
79 SB092	80 SB093	81 SB094	82 SB096	83 SB097	84 SB098	85 SB099	86 SB100	87 SB101	88 SB102	89 SB103	90 SB105	91 SB106
104 SB119	103 SB118	102 SB117	101 SB116	100 SB115	99 SB114	98 SB113	97 SB112	96 SB111	95 SB110	94 SB109	93 SB108	92 SB107
105 SB120	106 SB121	107 SB122	108 SB123	109 SB124	110 SB125	111 SB126	112 SB127	113 SB128	114 SB129	115 SB131	116 SB133	117 SB134
130 SB149	129 SB148	128 SB147	127 SB146	126 SB145	125 SB144	124 SB143	123 SB142	122 SB141	121 SB140	120 SB139	119 SB137	118 SB136
131 SB150	132 SB151	133 SB152	134 SB153	135 SB154	136 SB155	137 SB156	138 SB159	139 SB160	140 SB163	141 SB164	142 SB165	143 SB166
156 SB182	155 SB181	154 SB180	153 SB179	152 SB175	151 SB174	150 SB173	149 SB172	148 SB171	147 SB170	146 SB169	145 SB168	144 SB167
157 SB183	158 SB184	159 SB185	160 SB186	161 SB187	162 SB188	163 SB190	164 SB191	165 SB192	166 SB193	167 SB194	168 SERI	169 BABAX
182 SB001	181 SB014	180 SB029	179 SB047	178 SB064	177 SB079	176 SB092	175 SB107	174 SB120	173 SB136	172 SB150	171 SB167	170 SB183
183 SB184	184 SB168	185 SB151	186 SB137	187 SB121	188 SB108	189 SB093	190 SB080	191 SB065	192 SB048	193 SB030	194 SB016	195 SB002
208 SB003	207 SB018	206 SB031	205 SB049	204 SB066	203 SB081	202 SB094	201 SB109	200 SB122	199 SB139	198 SB152	197 SB169	196 SB185
209 SB186	210 SB170	211 SB153	212 SB140	213 SB123	214 SB110	215 SB096	216 SB082	217 SB067	218 SB050	219 SB032	220 SB019	221 SB004
234 SB005	233 SB020	232 SB036	231 SB052	230 SB068	229 SB083	228 SB097	227 SB111	226 SB124	225 SB141	224 SB154	223 SB171	222 SB187
235 SB188	236 SB172	237 SB155	238 SB142	239 SB125	240 SB112	241 SB098	242 SB084	243 SB070	244 SB053	245 SB038	246 SB021	247 SB006
260 SB007	259 SB022	258 SB039	257 SB054	256 SB071	255 SB085	254 SB099	253 SB113	252 SB126	251 SB143	250 SB156	249 SB173	248 SB190
261 SB191	262 SB174	263 SB159	264 SB144	265 SB127	266 SB114	267 SB100	268 SB086	269 SB072	270 SB055	271 SB040	272 SB023	273 SB008
286 SB009	285 SB024	284 SB041	283 SB057	282 SB073	281 SB087	280 SB101	279 SB115	278 SB128	277 SB145	276 SB160	275 SB175	274 SB192
287 SB193	288 SB179	289 SB163	290 SB146	291 SB129	292 SB116	293 SB102	294 SB088	295 SB074	296 SB058	297 SB043	298 SB025	299 SB010
312 SB011	311 SB026	310 SB044	309 SB061	308 SB076	307 SB089	306 SB103	305 SB117	304 SB131	303 SB147	302 SB164	301 SB180	300 SB194
313 SERI	314 SB181	315 SB165	316 SB148	317 SB133	318 SB118	319 SB105	320 SB090	321 SB077	322 SB062	323 SB045	324 SB027	325 SB012
338 SB013	337 SB028	336 SB046	335 SB063	334 SB078	333 SB091	332 SB106	331 SB119	330 SB134	329 SB149	328 SB166	327 SB182	326 BABAX

Figure 3.1: Field layout of `HEAT05` trial.

We will first do a design-based analysis of the HEAT05 trial. The Preliminary Single Environment Analysis (PSEA) window can be accessed from either:

- Stats | QTLs (Linkage/Association) | Phenotypic Analysis | Preliminary Single Environment Analysis; or,
- in the QTL Data View (Section 1.2) using shortcut Phenotypic analysis | Preliminary Single Environment Analysis (see Figure 3.2).
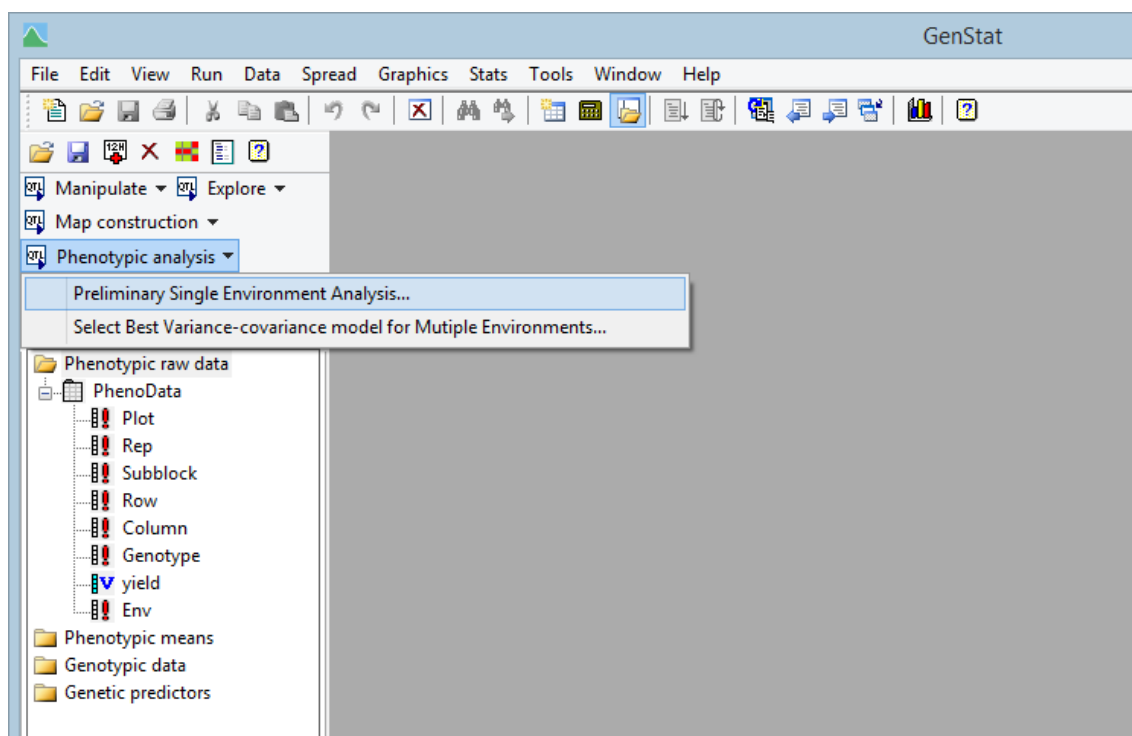
Figure 3.2: Accessing the PSEA menu from the QTL Data View.

For this example, select Incomplete block design from the Design: drop-down menu (there are a number of Design: options available that we will discuss in Section 3.5) and HEAT05 from the Environment: drop-down menu. Next select yield as the trait to be analysed by clicking in the Data: field and then double-clicking on yield from the Available data: field. Next select Genotype in the Genotypes: field, keep the default option of All genotypes in QTL analysis (the use of Extra genotypes present is described in Section 3.4). Specify the incomplete block model in Blocks: as Rep/Subblock using the Available data: and Operators: fields (Figure 3.3).
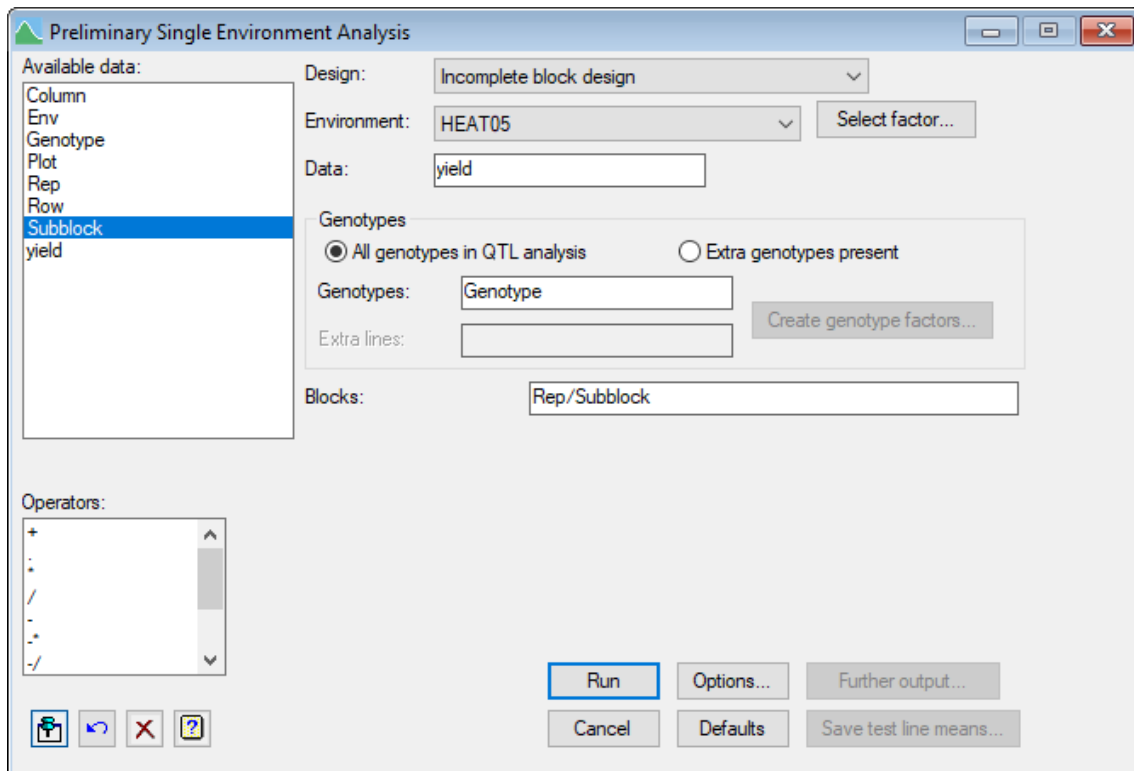
Figure 3.3: Preliminary Single Environment Analysis window, for Environment `HEAT05` from the `SB_yield.csv` data set.

The Preliminary Single Environment Analysis procedure performs two mixed model analyses: the first (Step 1) where the `Genotype` factor is fitted as a random term and the second (Step 2) where the `Genotype` factor is fitted as a fixed term: all other terms are as specified by the user in the menu. An overview of linear mixed models is given with some background theory in Chapter 10. The Step 1 model is used to obtain estimates of variance parameters. By default, these variance parameters are then used in the Step 2 model when the `Genotype` factor is fitted as a fixed term. The rationale for this process is that we would prefer to fit `Genotype` as a random term, as this avoids selection bias and results in better estimates of the variance parameters (Smith *et al.*, 2001), particularly for unreplicated designs with check plots, including augmented designs. However, the shrinkage associated with predictions of random effects (Section 10.5) is undesirable when predictions are to be carried forward to a second stage analysis (Smith *et al.*, 2001). For this reason, we use a compromise: we set `Genotype` as a fixed term in order to obtain unbiased estimates but use variance parameters estimated from the model with `Genotype` fitted as random. It is possible to override this process and re-estimate the variance

parameters in the second step (`Genotype` fixed) by checking Re-estimate variance parameters under the Analysis of Phenotypic Data Options menu (see Figure 3.4).

The default output from this process is a model summary, estimates of variance parameters and Wald tests (Section 10.6) for both steps. It is also possible to obtain an estimate of heritability for the trial from the Step 1 model, with `Genotype` fitted as random, by checking the Heritability box under Display on the Analysis of Phenotypic Data Options menu (Figure 3.4). This calculation is described in Section 3.3.
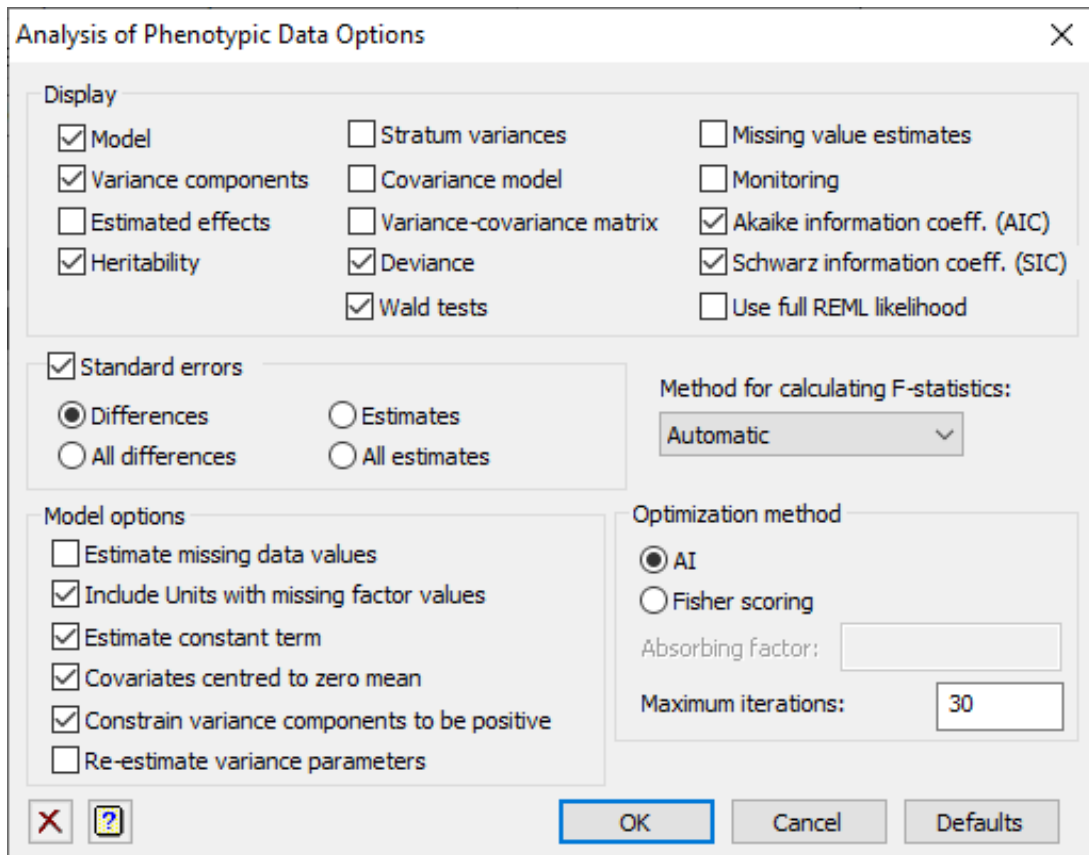
Figure 3.4: PSEA options.

Clicking on OK on the Analysis of Phenotypic Data Options menu returns focus to the PSEA menu, and clicking Run produces the following output for the `HEAT05` trial. Genstat commands have been omitted from the output, and bold font has been used to highlight the `Genotype` factor in the models.

```
REML variance components analysis
=================================

Response variate: yield_HEAT05
Fixed model:      Constant
Random model:     Genotype_HEAT05 + Rep_HEAT05 + Rep_HEAT05.Subblock_HEAT05
Number of units:   338

Residual term has been added to model

Sparse algorithm with AI optimisation


Estimated variance components
-----------------------------

Random term             component        s.e.
Genotype_HEAT05            1362.9        176.4
Rep_HEAT05                 1216.0       1742.7
Rep_HEAT05.Subblock_HEAT05
                            178.3         68.5


Residual variance model
-----------------------

Term                    Model(order)  Parameter      Estimate      s.e.
Residual                Identity      Sigma2            429.4      50.4


Deviance: -2*Log-Likelihood
---------------------------

                  Deviance    d.f.
                   2760.76   333

Note: deviance omits constants which depend on fixed model fitted.


      Akaike information coefficient      2768.76
Schwarz Bayes information coefficient     2784.04

Note: omits constants, (n-p)log(2pi) - log(det(X'X)), that depend only on the
fixed model.

(based on the residual log-likelihood)


Heritability:      0.8528
```

56

```
1406...................................................................

REML variance components analysis
=================================

Response variate:   yield_HEAT05
Fixed model:        Constant + Genotype_HEAT05
Random model:       Rep_HEAT05 + Rep_HEAT05.Subblock_HEAT05
Number of units:    338

Residual term has been added to model

Sparse algorithm with AI optimisation

Estimated variance components
-----------------------------

Random term               component        s.e.
Rep_HEAT05                  1214.9         fixed
Rep_HEAT05.Subblock_HEAT05
                            178.1          fixed

Residual variance model
-----------------------

Term                      Model(order)  Parameter      Estimate      s.e.
Residual                  Identity      Sigma2          429.0        46.7

Deviance: -2*Log-Likelihood
---------------------------

                   Deviance   d.f.
                    1348.11   168

Note: deviance omits constants which depend on fixed model fitted.

Tests for fixed effects
-----------------------

Sequentially adding terms to fixed model

Fixed term                Wald statistic  n.d.f.  F statistic  d.d.f.    F pr
Genotype_HEAT05                 1167.88    168          6.95   169.0  <0.001

Dropping individual terms from full fixed model

Fixed term                Wald statistic  n.d.f.  F statistic  d.d.f.    F pr
Genotype_HEAT05                 1167.88    168          6.95   169.0  <0.001
```

```
* MESSAGE: denominator degrees of freedom for approximate F-tests are calculated
using  algebraic  derivatives  ignoring  fixed/boundary/singular  variance
parameters.



      Akaike information coefficient      1350.11
Schwarz Bayes information coefficient      1353.24

Note: omits constants, (n-p)log(2pi) - log(det(X'X)), that depend only on the
fixed model.

(based on the residual log-likelihood)
```

There are several things to note here. Because the data set contains yields for several environments, the data has been subsetted to create new structures corresponding to the environment to be analysed. The names of these new structures are constructed by appending the environment name to the ends of the original structure names. Thus `yield_HEAT05` indicates that this variate contains yield values for trial `HEAT05` only. The model summary indicates the terms that have been fitted. In the first model, the `Genotype_HEAT05` factor is fitted as a random term, with the terms corresponding to the blocking structure for trial `HEAT05` (`Rep_HEAT05` and `Rep_HEAT05.Subblock_HEAT05`) also fitted as random terms. An overall constant term (`Constant`) is automatically added into the fixed model. The estimated variance components are printed after the model summary. In the first model, it is clear that genotypes account for a substantial proportion of the variation. This is quantified by the estimate of heritability as 0.85 - this quantity is discussed further in Section 3.3. In the second model, the factor `Genotype_HEAT05` has been switched from the random to the fixed model, and appears in the table of tests for fixed terms (see Section 10.6). Again, the approximate F-test suggests there is very strong evidence for differences among the set of genotypes.

Before proceeding, it is sensible to check whether the assumptions underlying the analysis have been met (see Section 10.7) by checking the residuals. Residual plots can be generated by using the Further output button following analysis, which generates the Single Environment REML Analysis Further Output window shown in Figure 3.5. Clicking on Residual plots brings up the window on the right-hand side of Figure 3.5.
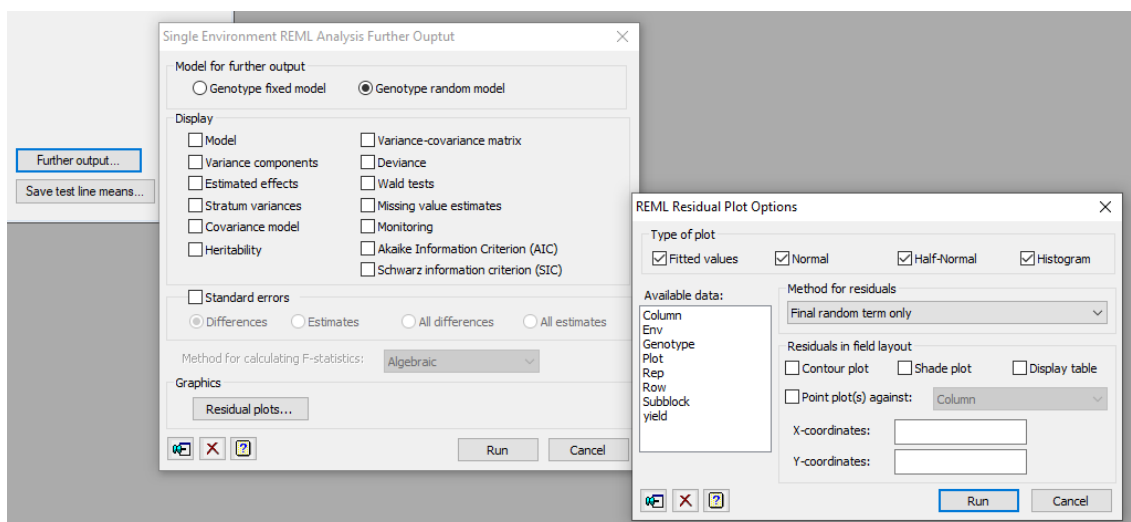
Figure 3.5: Further output and residual plot menu options for PSEA.

The default output is a composite set of residual plots, based on the first step analysis, with `Genotype` fitted as a random term. Residuals from the second step analysis (`Genotype` fixed), as shown in Figure 3.6, are obtained by selecting Genotype fixed model on the Single Environment REML Analysis Further Output window (left-hand side of Figure 3.5). There are several different definitions of residuals from the linear mixed model (see Section 10.7). The Method for residuals drop-down allows five choices for these plots: residuals may be constructed from all random terms (setting Combine all random terms), corresponding to the marginal residuals of Section 10.7. Alternatively, residuals may be constructed from the final random term only (the model deviations) corresponding to the conditional residuals of Section 10.7 (Final random term only). Standardized marginal and conditional residuals can be constructed by setting the Method for residuals to Standardized residuals from all random terms or Standardized residuals from final random term only, respectively. Marginal residuals can also be constructed from all random terms except spline terms (Combine all random terms, excluding spline terms).

In Figure 3.6 conditional residuals are plotted setting Method for residuals to Final random term only. As described in Section 10.7, for this model, the residuals should be consistent with an independent sample from a Normal distribution with constant variance. In Figure 3.6, the histogram of residuals is reasonably symmetric, there is no evidence of variance changing in relation to the fitted values and the Normal plots are reasonably close to a straight line, giving no evidence of departures from the model assumptions of normality and constant variance.
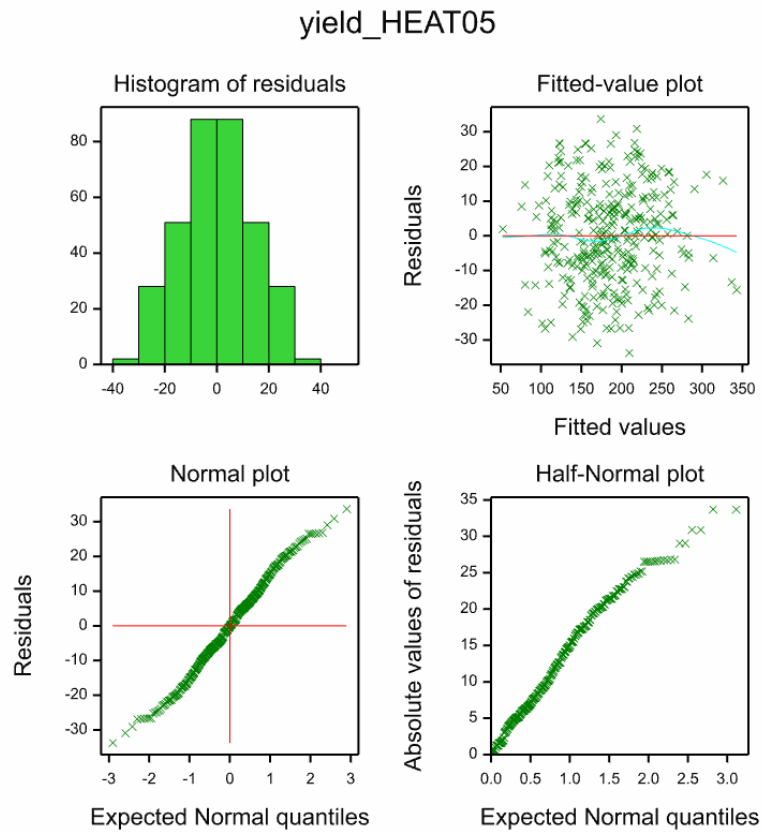
Figure 3.6: Conditional residual plots for `HEAT05` trial with `Genotype` fitted as fixed term.

However, these residual plots cannot give information about dependence between residuals. To assess independence we can look at plots of residuals in field layout (i.e. shade or contour plots), as well as the variograms (described in Section 3.6). Figure 3.7 shows specification of a shade plot.

Shade plots for both marginal and conditional residuals (Method for residuals settings Combine all random terms and Final random term only, respectively) for trial `HEAT05`, based on the second step analysis (`Genotype` fixed) are shown in Figure 8. The left-hand plot, based on the combined residuals, shows the large effect of replicate, as residuals in the first replicate (rows 1-13) are clearly smaller than those in the second replicate (rows 14-26). Once the replicate and sub-block effects have been removed (right-hand plot), some left to right trend is visible across the columns.
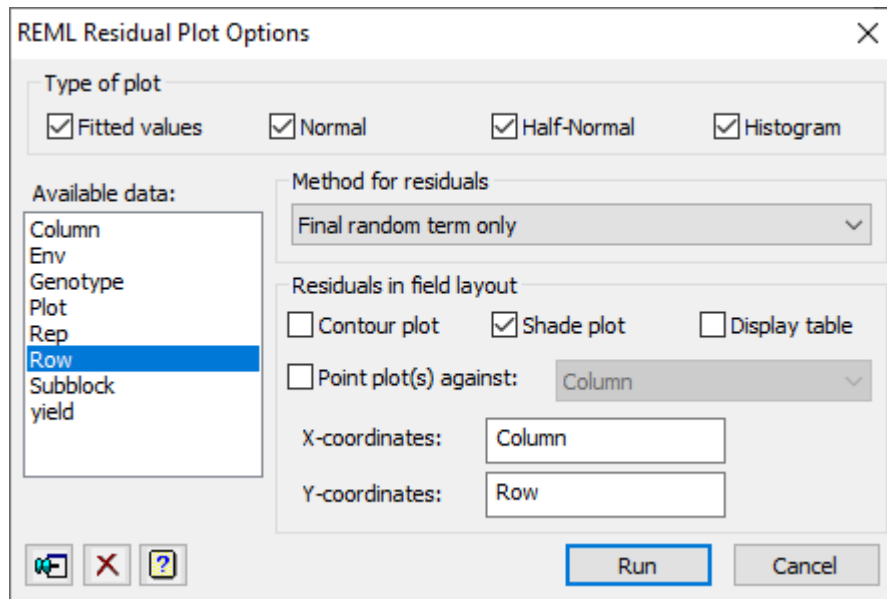
Figure 3.7: Specifying a shade plot of residuals in field layout.
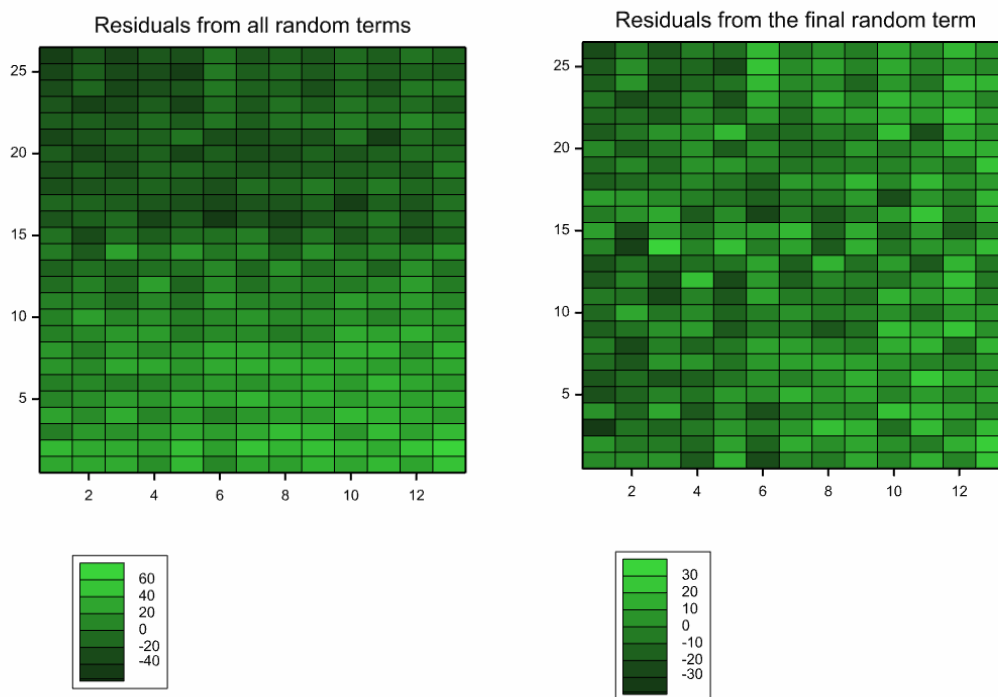
Figure 3.8: Shade plots of residuals in field layout for `HEAT05` trial.

In practice, we might investigate the residuals more closely as part of the trial analysis, using the exploratory techniques illustrated in Section 3.6 prior to entering the QTL system.

3.2   Generating trait means

Once a satisfactory model has been fitted, the trait means are predicted from the second step performed by the Preliminary Single Environment Analysis menu, that is, when `Genotype` is fitted as a fixed term. These means are called the Best Linear Unbiased Estimates (BLUEs) (see Section 10.6). The BLUEs can be saved using the Save test line means button (Figure 3.3). The Save Test Line Means window specifies the data structures to be saved (Figure 3.9).

The Environment label: is used to identify the results from an analysis for a particular environment (e.g. `HEAT05`).
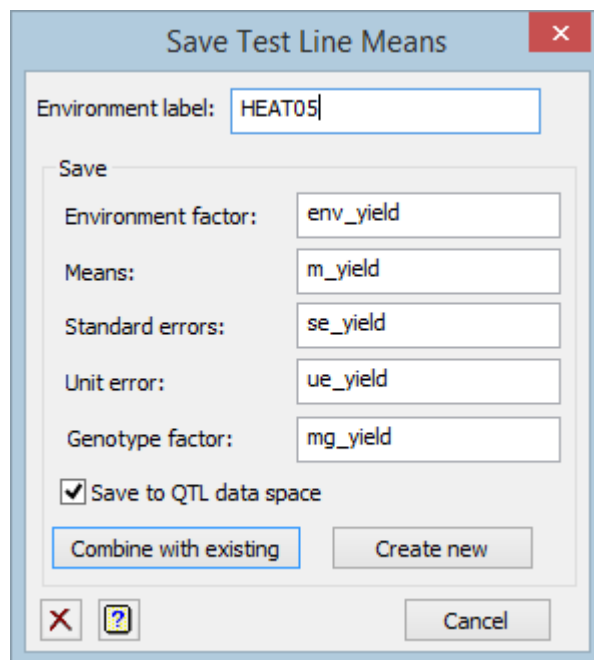
Figure 3.9: Save Test Line Means window.

Default names are suggested for the structures to be saved (Figure 3.9). These structures correspond to the trait means for each genotype (or line) and so have a different

62

length to the original variables which relate to the raw data - new names are therefore required to prevent the original variables being overwritten. The default names generated use the trait name as a suffix (e.g. `_trait`). The quantities saved are the environment and genotype labels (default names `env_trait` and `mg_trait`), the predicted means and their standard errors (`m_trait` and `se_trait`) and a set of unit errors (`ue_trait`).

The unit errors (`ue_trait`) are a measure of precision of the trait means suitable for use as weights in G×E and QTL×E analyses, and their inclusion can improve genotype predictions when trials within a data set have different levels of precision (Welham *et al.*, 2010; Möhring & Piepho, 2009; see also Section 4.3). The unit errors are derived from the estimated variance-covariance matrix of genotype predictions from the Step 2 model (`Genotype` fitted as fixed), denoted $\hat{\Sigma}$. For *n* genotypes, let $\Pi = (\pi_1, \ldots, \pi_n)'$ denote a vector of weights obtained by taking the diagonal elements of $\hat{\Sigma}^{-1}$, i.e. the diagonal of the inverse of this variance-covariance matrix. The unit errors are calculated as the inverse of these weights, i.e. $1/\pi_i$ for *i=1…n*, as suggested by Smith *et al.* (2001).

The structures generated can be saved in two different ways: as a new set of structures (Create new) or used to update an existing set in the QTL Data Space (Combine with existing):

- Create new will overwrite any other structures of the same name. It can be used to create the structures for the first trial analysed in a multi-environment data set or to re-start the generation of the means.

- Combine with existing will append the new values onto the named structures; or, if the named structures do not exist, it will create the new structures. In this way, a composite data set (consisting of trait means, SEs, unit errors and factors giving genotype and environment labels) can be formed for G×E and QTL×E analysis from repeated use of the Preliminary Single Environment Analysis menu, with Save Test Line Means | Combine with existing to build up a set of trait means for all environments in the data set.

By default these new variables will be added into the QTL Data Space.

Some caution is required if means are generated for more than one trait (as required for multi-trait analysis), as the QTL Data Space can currently only hold one indexing factor for both genotypes and environments. If some traits are measured in different subsets of environments, then different indexing factors are required - these can be constructed using the Save Test Line Means menu and will be stored within Genstat, but only the last set generated will be present in the QTL Data Space and appear automatically on menus.

3.3   Calculating heritability

Heritability is commonly used by plant breeders to quantify the degree of genetic determination of the trait of interest, and in the simplest case can be interpreted as the proportion of observed variation that can be attributed to genetic differences (see Kearsey & Pooni, 1996). Heritability (denoted h^2) ranges from 0 (no genetic determination of the trait) to 1 (total genetic determination with no measurement error). Heritability depends on both the trait measured and the precision of the trial. In addition, there are many different definitions of heritability, depending on whether it relates to total genetic variation (broad-sense) or additive genetic effects (narrow-sense) or to genotype predictions (mean line) or individual observations. Because heritability relates to genetic variation, it can only be obtained when `Genotype` is fitted as random (i.e. the first analysis performed by the PSEA menu). For simple variance component models, heritability can be obtained directly from the estimated variance components. For more complex models, such as spatial models (Section 3.6), we require a more general definition. We therefore use the generalized heritability measure described by Cullis, Smith & Coombes (2006), defined as

$$h^2 = 1 - \frac{\bar{v}_{BLUP}}{2\hat{\sigma}_g^2}$$

where $\bar{v}_{BLUP}$ is the mean variance of a difference between two genotype BLUPs (see Section 10.6 for background information on BLUPs) and $\hat{\sigma}_g^2$ is the estimated variance component for `Genotype`. This quantity can be interpreted as a broad-sense mean line heritability, derived from an estimate of the correlation between the genotype BLUPs and their unknown true value. The estimated trial heritability can be displayed by checking the Heritability box under Display on the PSEA Options menu (see Figure 3.4).

3.4   Modelling the genotype structure: test and extra lines

Many trials include extra (often control, check, standard or parental) genotype lines, as well as lines from a mapping (or more general) population. It is almost always better to analyse the whole trial (i.e. using the full set of lines present) to obtain predicted trait means to take forward to a QTL analysis rather than analyse a portion of the trial. Analysis of the full trial gives the best estimate of variance parameters that will be used to combine information across strata (for unbalanced designs) and estimates of uncertainty in predictions. Analysing portions of a trial is undesirable, as it can lead to different (and

usually less reliable) estimates of the variance parameters. The Genotypes: section of the Preliminary Single Environment Analysis window (Figure 3.3) allows the `Genotype` factor to be partitioned so that the whole trial is analysed but some genotypes can be excluded from estimates of heritability and their means not saved for QTL analysis.

In the simplest case, where all genotypes in the trial are to be included in the QTL analysis, the default options are adequate: select the All genotypes in QTL analysis option and specify the name of the factor in the Genotypes: input field. If some genotypes present in the trial, such as check lines or parents, are to be excluded from the QTL analysis, then select the Extra genotypes present option. In this case, two sets of genotypes need to be defined: the test lines, genotypes to be taken forward for QTL analysis, and the extra lines, genotypes to be excluded from QTL analysis. These two sets can be generated from a single factor by using the Create genotype factors button (see Figure 3.10).
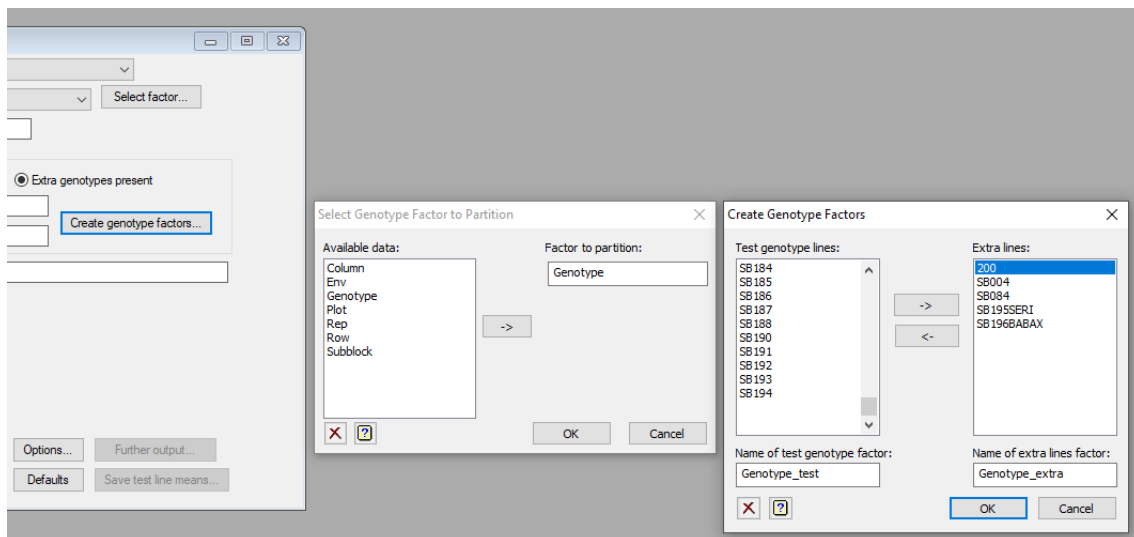
Figure 3.10: Defining the `Genotype` structure to accommodate 'extra' genotype lines such as standards, controls, or parental lines.

First, select the `Genotype` factor to partition. This brings up the Create Genotype Factors window, with the full list of factor labels shown in the Test genotype lines: box. Select those lines to be excluded from QTL analysis and move these to the Extra lines: box. Each set of labels is used to form a new factor, by default called `Genotype_test` and `Genotype_extra`, respectively. The `Genotype_test` factor has levels/labels corresponding to the labels in the Test genotype lines: box, with missing values (blank cell) used to represent the remaining (extra) lines. The `Genotype_extra` factor has levels/labels corresponding to the labels in the Extra lines: box, plus an additional

level/label (e.g. level 0 and/or label 'Test line') used to identify the set of genotypes taken forward for QTL analysis.

For example, for the CIMMYT spring wheat trials data set there were 167 lines from a mapping population (labels `SBxxx`), two parents (labels `SB195SERI` and `SB196BABAX`), and a variety used to fill out the design at one site (label `200`). The parents and the filler variety are not required for QTL analysis, and two lines (`SB004` and `SB084`) from the mapping population cannot be included as they have no genotypic data - these five labels are moved into the Extra lines: box (Figure 3.10). The labels of the newly created `Genotype_test` and `Genotype_extra` are shown in Table 3-1.

Table 3-1: Partitioning of `Genotype` factor for `HEAT05` trial into test lines (to be taken forward to QTL analysis) and extra lines (to be excluded from QTL analysis).

Full set	Test lines	Extra lines
200		200
SB001	SB001	Test line
SB002	SB002	Test line
SB003	SB003	Test line
SB004		SB004
SB005	SB005	Test line
⋮	⋮	⋮
SB083	SB083	Test line
SB084		SB084
SB085	SB085	Test line
⋮	⋮	⋮
SB193	SB193	Test line
SB194	SB194	Test line
SB195SERI		SB195SERI
SB196BABAX		SB196BABAX

The `Genotype_extra` factor is fitted as a fixed term in both steps of the analysis, whereas the `Genotype_test` factor will be used as a random term in the first step and as a fixed term in the second step. The output from this analysis for the `HEAT05` trial is shown below:

```
REML variance components analysis
=================================


Response variate:   yield_HEAT05
Fixed model:        Constant + Genotype_extra_HEAT05
Random model:       Genotype_test_HEAT05 + Rep_HEAT05 + Rep_HEAT05.
Subblock_HEAT05
Number of units:    338


Residual term has been added to model


Sparse algorithm with AI optimisation
Units with missing factor/covariate values included
   - specific effect for term(s) omitted for units with missing values in
Genotype_test_HEAT05



Estimated variance components
-----------------------------


Random term              component       s.e.
Genotype_test_HEAT05       1390.4        181.6
Rep_HEAT05                 1215.9       1742.7
Rep_HEAT05.Subblock_HEAT05
                            180.3         69.2



Residual variance model
-----------------------


Term                     Model(order)  Parameter       Estimate      s.e.
Residual                 Identity      Sigma2            429.1        50.3



Deviance: -2*Log-Likelihood
---------------------------


                  Deviance   d.f.
                  2730.09    329

Note: deviance omits constants which depend on fixed model fitted.



Tests for fixed effects
----------------------


Sequentially adding terms to fixed model

Fixed term               Wald statistic  n.d.f.   F statistic  d.d.f.   F pr
Genotype_extra_HEAT05              7.13       4          1.78   166.1  0.135
```

```
Dropping individual terms from full fixed model

Fixed term              Wald statistic  n.d.f.   F statistic  d.d.f.    F pr
Genotype_extra_HEAT05            7.13       4          1.78   166.1   0.135

* MESSAGE: denominator degrees of freedom for approximate F-tests are calculated
using   algebraic   derivatives   ignoring   fixed/boundary/singular   variance
parameters.


        Akaike information coefficient     2738.09
Schwarz Bayes information coefficient     2753.32

Note: omits constants, (n-p)log(2pi) - log(det(X'X)), that depend only on the
fixed model.

(based on the residual log-likelihood)


Heritability:        0.8553

565....................................................................


REML variance components analysis
=================================

Response variate:  yield_HEAT05
Fixed model:       Constant + Genotype_extra_HEAT05 + Genotype_test_HEAT05
Random model:      Rep_HEAT05 + Rep_HEAT05.Subblock_HEAT05
Number of units:   338


Residual term has been added to model


Sparse algorithm with AI optimisation
Units with missing factor/covariate values included
   - specific effect for term(s) omitted for units with missing values in
Genotype_test_HEAT05


Estimated variance components
----------------------------

Random term            component        s.e.
Rep_HEAT05                  1214.2       fixed
Rep_HEAT05.Subblock_HEAT05
                             180.1       fixed


Residual variance model
----------------------
```

```
Term                           Model(order)  Parameter      Estimate      s.e.
Residual                       Identity      Sigma2            428.5      46.6



Deviance: -2*Log-Likelihood
---------------------------

                   Deviance   d.f.
                    1348.11   168

Note: deviance omits constants which depend on fixed model fitted.



Tests for fixed effects
-----------------------

Sequentially adding terms to fixed model

Fixed term               Wald statistic  n.d.f.   F statistic  d.d.f.    F pr
Genotype_extra_HEAT05             10.57       4          2.64   169.0   0.035
Genotype_test_HEAT05           1158.63     164          7.06   169.0  <0.001

Dropping individual terms from full fixed model

Fixed term               Wald statistic  n.d.f.   F statistic  d.d.f.    F pr
Genotype_test_HEAT05           1158.63     164          7.06   169.0  <0.001

* MESSAGE: denominator degrees of freedom for approximate F-tests are calculated
using  algebraic  derivatives  ignoring  fixed/boundary/singular  variance
parameters.


      Akaike information coefficient      1350.11
Schwarz Bayes information coefficient      1353.24

Note: omits constants, (n-p)log(2pi) - log(det(X'X)), that depend only on the
fixed model.

(based on the residual log-likelihood)
```

When you use the Save Test Line Means menu for this analysis (Figure 3.9), only the genotypes specified in `Genotype_test` are estimated and stored for subsequent analyses. Similarly, the genotype variance component and the heritability are estimated only for those genotypes specified in `Genotype_test`, and so may differ from the previous estimate, particularly if the excluded lines are very different from those retained for QTL analysis. The other variance parameter estimates may also change, although usually these changes will be small.

69

3.5 Trial design and variance models

The Preliminary Single Environment Analysis menu provides a choice of trial designs (via the Design: box, see Figure 3.3), and incorporates a subset of the facilities for model investigation available in the Mixed Models (REML) menu (Stats | Mixed Models (REML)). The appearance of the PSEA menu alters according to the type of experimental design selected, giving options appropriate to that design. The pre-defined designs encompass some of the most popular choices for plant breeding trials, but a more general model can also be specified. In this section we will describe the different designs available, and some common models. The process of model selection is described in Section 3.6.

3.5.1 Randomized complete block design

The randomized complete block design (RCBD) comprises a number of blocks (sometimes called replicates), each containing all of the genotype lines in the trial, with genotypes allocated to plots within blocks completely at random. This design is appropriate where plots within blocks can be considered as homogeneous, but differences are expected between blocks. Figure 3.11 shows the layout of the PSEA menu if Design: option Randomized complete block design is chosen. The blocking structure should be specified in the Blocks: box, e.g. as `Block/Plot` where factor `Block` defines which block each plot belongs to, and factor `Plot` labels plots within each block.
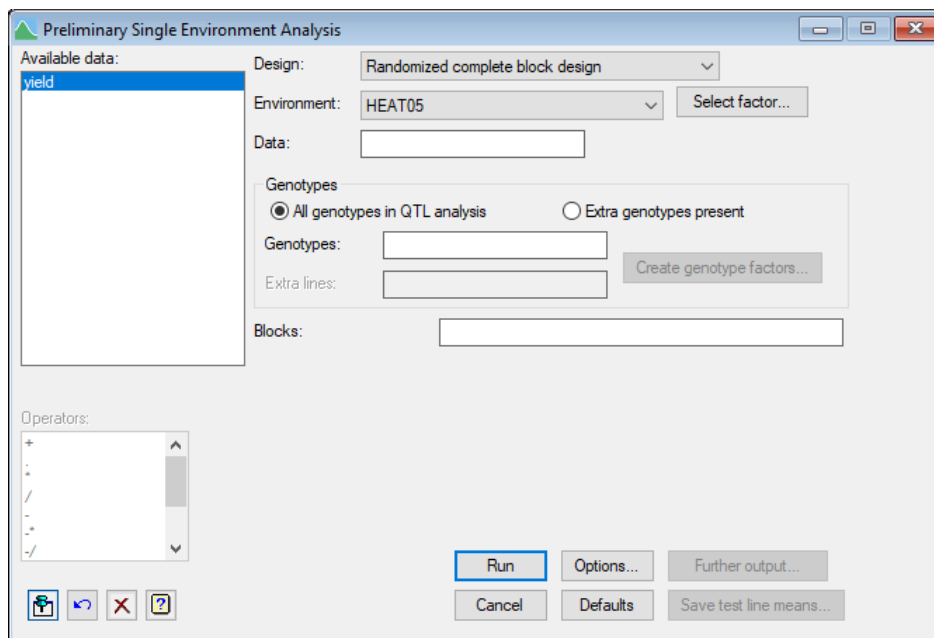
Figure 3.11: Layout of the PSEA menu for a randomized complete block design.

3.5.2 Incomplete block design

The randomized incomplete block design comprises a number of replicates which each contain all of the genotype lines in the trial. Each replicate is split into a number of blocks, usually of equal sizes, but the number of units per block is less than the number of genotypes. Each block therefore contains a subset of the genotypes. There are many different types of incomplete block design (BIBD, lattice design, alpha design, cyclic design) which aim to maximize the precision of genotype comparisons by balancing the co-occurrence of pairs of genotypes within the same block. This design is appropriate where replicates containing the full set of genotypes cannot be considered homogeneous, but smaller blocks within replicates can be considered homogeneous. The CIMMYT spring wheat trials used incomplete block designs, as described in Section 1.3.3 and shown in Figure 3.1. Figure 3.3 showed the layout of the PSEA menu if Design: option Incomplete complete block design is chosen. The blocking structure, e.g. `Rep/Subblock` for the `HEAT05` trial, is specified using the Blocks: box. Note that if plots had been labelled within sub-blocks as 1-13 by factor `SPlot`, then the structure could have been fully specified as `Rep/Subblock/SPlot`.

3.5.3 Spatial design in regular grid

The structure of an incomplete block design implies correlation between measurements on plots that depends on their co-location within the same replicate or sub-block (see Section 10.2). This correlation structure is discrete, with step changes at replicate and sub-block boundaries. In principle, we might expect the pattern of correlation to depend more on physical spatial proximity than on a pre-defined blocking structure. An approach based on modelling observed patterns of correlation within an experiment would therefore use random effects with correlation based on spatial location. However, ignoring the experimental design is unsatisfactory, as the randomization structure generates the correct strata and degrees of freedom (df) for testing fixed model terms. This is a particularly important consideration for designs where treatments are applied at a level higher than the observational unit; for example, split plot designs, where multiple samples are taken from a plot, or when technical replicates are used to improve the measurement process. We therefore recommend a hybrid approach, where we use random terms to account for the structure of the experimental design and add terms as required to take account of additional patterns of correlation.

For field trials laid out on a grid pattern, with regular spacing of rows and columns, the most common patterns of observed correlation comprise trends across the trial, row

and/or column effects which may be associated with field management operations, and spatial correlations across rows and/or columns (see Gilmour *et al.*, 1997). All of these options are enabled by using the Design: setting Spatial design in regular grid (Figure 3.12). The different model components and the question of model selection are discussed in Section 3.6.
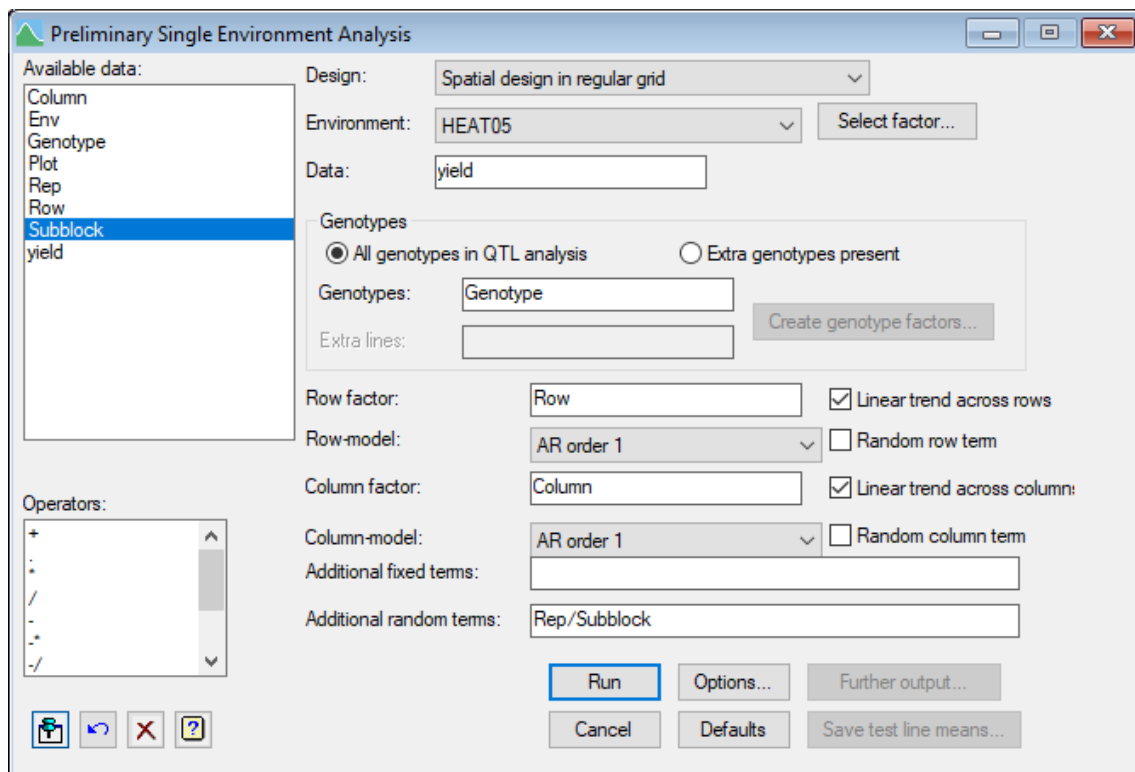
Figure 3.12: Layout of the PSEA menu for a spatial design in regular grid for HEAT05 trial.

Spatial correlation is usually imposed at the plot level for a field trial, which corresponds to the model deviations or residual term unless the measurements used multiple samples per plot or technical replication. It is common to assume that correlation across rows and columns act independently, which leads to a separable correlation structure. The auto-regressive structure of order 1 (AR1) is a flexible model for serial correlation that is often used in this context. For example, a model for spatial correlation in the HEAT05 trial might use a separable AR1 process across both rows and columns, applied at the plot level. Then if plot i is in row r_i and column c_i, and plot j is in row r_j and column c_j, the spatial covariance between these two plots is modelled as

$$\text{cov}(e_i, e_j) = \sigma^2 \rho_r^{|r_i - r_j|} \rho_c^{|c_i - c_j|}.$$

This spatial covariance depends on the residual variance, σ^2, the spatial correlation across rows, ρ_r, the spatial correlation across columns, ρ_c, and the displacement between the plots in the row and column directions. Covariance due to other random terms is added onto this quantity. The correlation parameters, ρ_r and ρ_c, must lie between -1 and 1 and so correlation decreases as the row and column displacement increases (see *Genstat Statistics Guide, Section 5.4,* for more details). This model can be simplified by allowing one of the spatial processes to become an identity model, in which case no spatial correlation is generated in that direction by this term.

The PSEA menu for Design: setting Spatial design in regular grid deals with the case of a regular rectangular grid, with units indexed by factors for rows and columns and all row and column combinations present. If data for some plots is missing, then the plots should be included in the data set, with their row and column values set, but with a missing value for the trait value(s). For this menu, only one measurement per plot is allowed. If there are several measurements per plot which are true replicates, e.g. sub-samples or technical replicates, and if there are the same number of measurements in each plot, then it is valid to analyse plot means. In all other cases, or for a non-rectangular layout, it will be necessary to analyse the data using the appropriate menu under Stats | Mixed models (REML), then construct the set of predicted trait means as described in Section 3.5.5 below.

Figure 3.12 shows the layout of the PSEA menu with Spatial design in regular grid, using settings for the final `HEAT05` model established in Section 3.6. This model uses a separable AR1 structure across rows and columns, in addition to the design blocking structure (`Rep/Subblock`) and added linear trend across both rows and columns. Explanation of this model is given in Section 3.6.

For this Design: setting, it is possible to obtain plots of two-dimensional variograms from the Further output button in order to diagnose problems with the spatial model, as discussed in Section 3.6.

3.5.4 General designs

The General design setting is intended to cope with other forms of experimental designs, or variations on the standard models. The layout for the PSEA menu using this Design: setting is shown in Figure 3.13. Additional fixed (Additional fixed terms:) and random (Additional random terms:) terms can be added to the baseline model consisting of the `Genotype` factor.
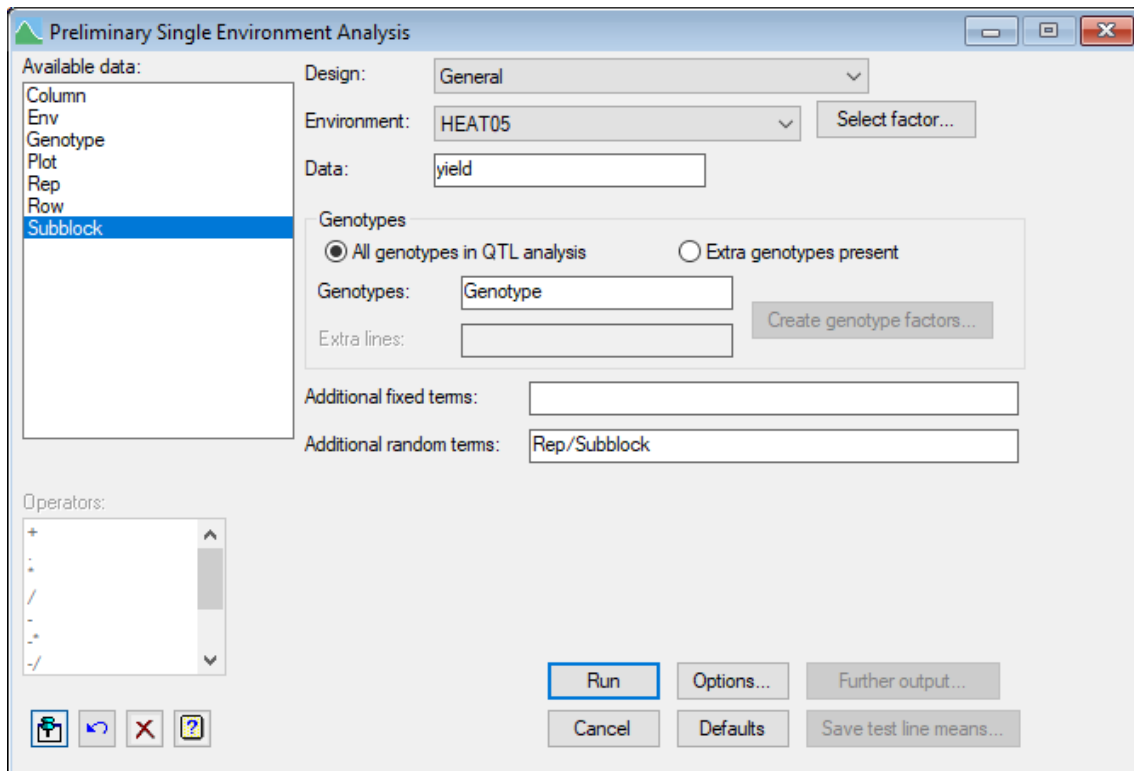
Figure 3.13: Layout of the PSEA menu for a general design.

As previously, the random terms will usually comprise the randomization and other physical structure of the design. However, it may occasionally be helpful to fit some of the blocking structure as fixed terms, and this menu makes that possible. Usually, only structure at the top level of a design, corresponding to complete replicates (and therefore containing no treatment information) and with few levels, would be fitted as fixed (e.g. replicates in an incomplete block design). The rationale for this approach is that estimation of variance components can be very poor for terms with few (<5) levels and the estimate may be negative, especially if differences at that level are not large. Even if there is some treatment information at this level, it may not be sensible to use the poorly estimated variance component as a basis for combination of information, and fitting a block term as fixed means that only treatment information from lower strata will be used.

3.5.5 Other designs

The Design: settings on the PSEA menu cannot deal with the full range of experimental designs that may be encountered in practice, and an alternative strategy must be followed for these other types of design. This may be necessary for spatial models with an irregular

layout, or with multiple measurements per plot, models for repeated measurements, or models including spline terms. In these cases, it is necessary to first identify a suitable model for the trait in each environment - some guidance on procedure is given for spatial models in Section 3.6 below, and a general recipe for building mixed models is given in Section 10.8. Further information on the various options given under Stats | Mixed models (REML) can be found in *A Guide to REML in Genstat*. Once a model has been identified, then predictions of trait means can be made using the Predict button. Alongside the predictions, it is possible to save standard errors and the full variance-covariance matrix of the predictions, from which unit errors can be derived as described in Section 3.2. These predictions, SEs and unit errors can be stored in a Genstat spreadsheet with the genotype labels, and a text column giving the environment name added. Once several spreadsheets have been accumulated for the different environments, these can be appended together (Spread | Manipulate | Append, see Figure 3.14) to give a data set of trait means that can be imported directly into the QTL Data Space.
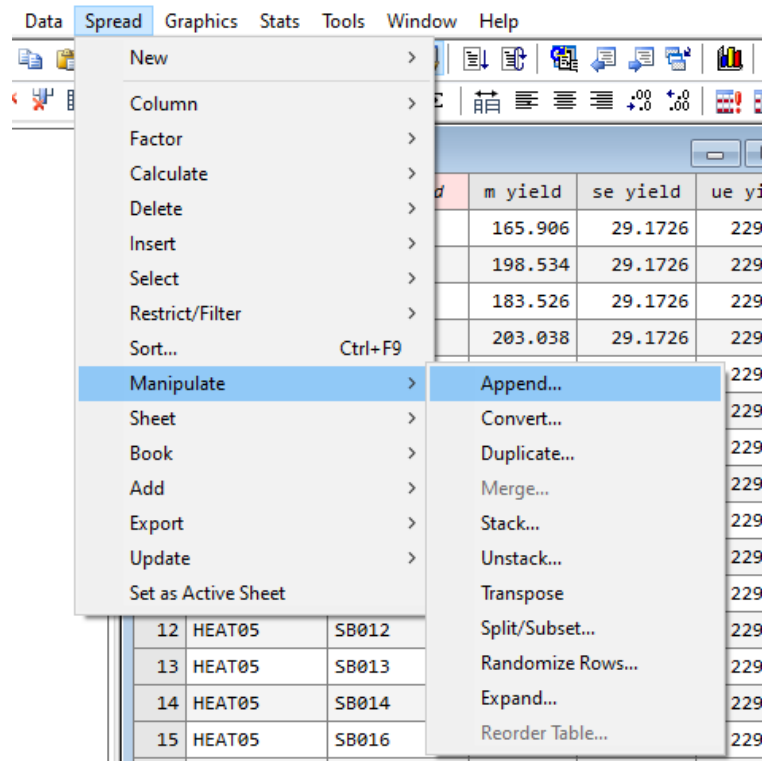
Figure 3.14: Appending Genstat spreadsheets.

3.6   Variance modelling

In this section, we illustrate variance modelling via the specific example of spatial trends in the `HEAT05` field trial. Examples in different contexts (variance components model, longitudinal data) can be found in *A Guide to REML in Genstat* or the *Genstat Statistics Guide*.

Some field trials for mapping populations are large (e.g. > 500 genotypes) and can take up large areas of land. This increases the chance that spatial trends will be present in the field. In addition, crop management operations tend to act along columns and/or rows of a trial, and can induce additional (extraneous) patterns of variation. These natural and extraneous spatial trends can be accounted for using spatial modelling. More details on this approach can be found, with examples, in Gilmour *et al.* (1997) and Stefanova *et al.* (2009). Here we illustrate the approach using the `HEAT05` field trial, loaded as the individual trial from Genstat workbook `SB_HEAT05.gwb`. The field layout (shown in Figure 3.1) consisted of a rectangular array with 26 rows and 13 columns. The two replicates corresponded to rows 1-13 and rows 14-26, and the sub-blocks were the individual rows. This corresponds to a regular grid, so we will work with the menu for building spatial mixed models (Stats | Mixed Models (REML) | Spatial Models | Regular Grid, see Figure 3.15).
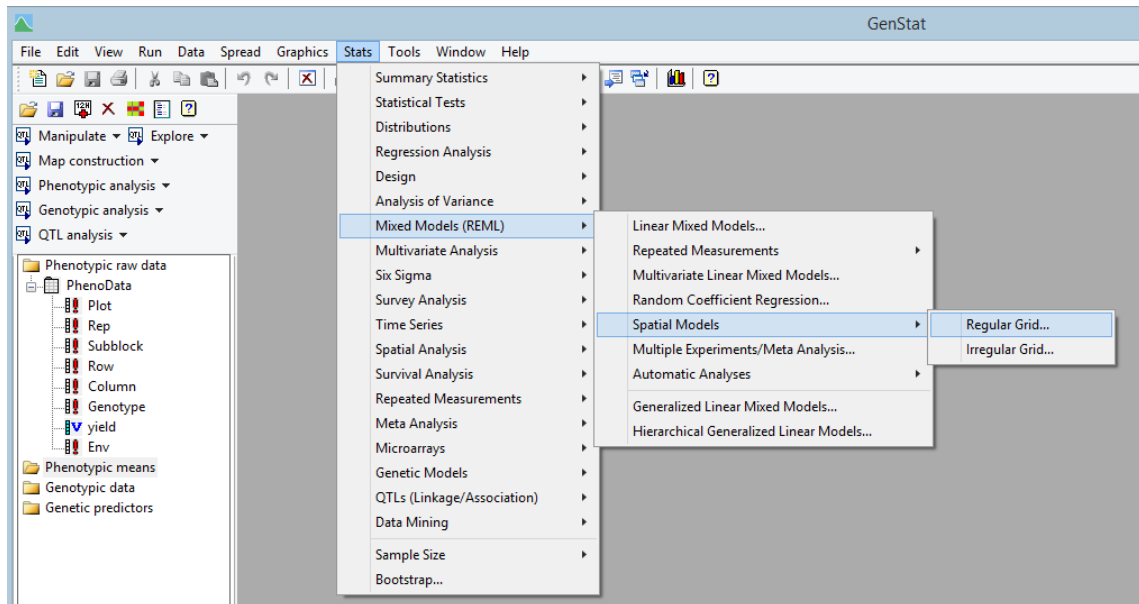
Figure 3.15: Menus for spatial analysis using mixed models.

In Section 10.8, we give a recipe for building a mixed model, and we will broadly follow that approach. However, we will add a preliminary stage where we investigate the spatial trend, and we will do this ignoring the structure of the experimental design. For the reasons given in Section 10.2, we would not normally consider a model excluding these terms because of their potential importance in testing fixed terms, but we exclude them for this exploratory step in order to simplify interpretation of the initial diagnostic plots. As our exploratory model, we fit a separable AR1×AR1 model across rows and columns, specified as shown in Figure 3.16. The `Row` factor labels the rows of the layout and the `Column` factor labels the columns. The factor `Genotype` is included as a random term.
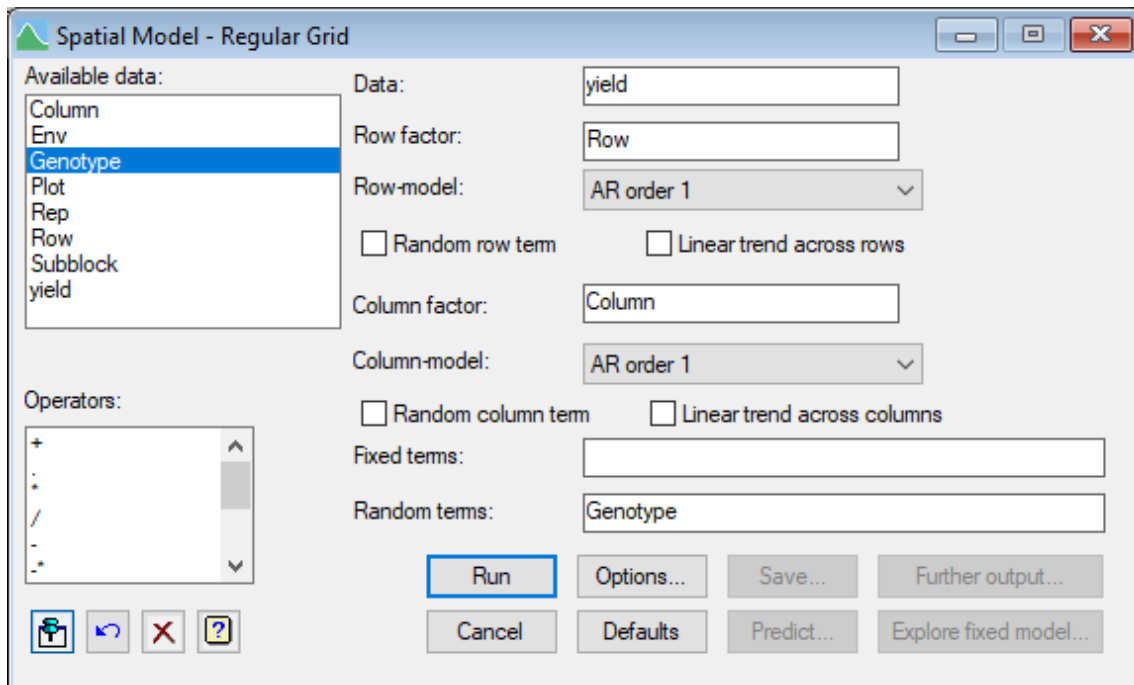
Figure 3.16: Specification of exploratory model for `HEAT05` trial.

We will use the residuals from this model to investigate the presence of spatial trend. After the model has been fitted (by clicking on the Run button), the Save button becomes active and can be used to save residuals and various other results from the analysis (Figure 3.17). In Figure 3.17, we have chosen to form residuals from the final random term only (see Section 10.7), and to save the residuals in a structure called `Res`.
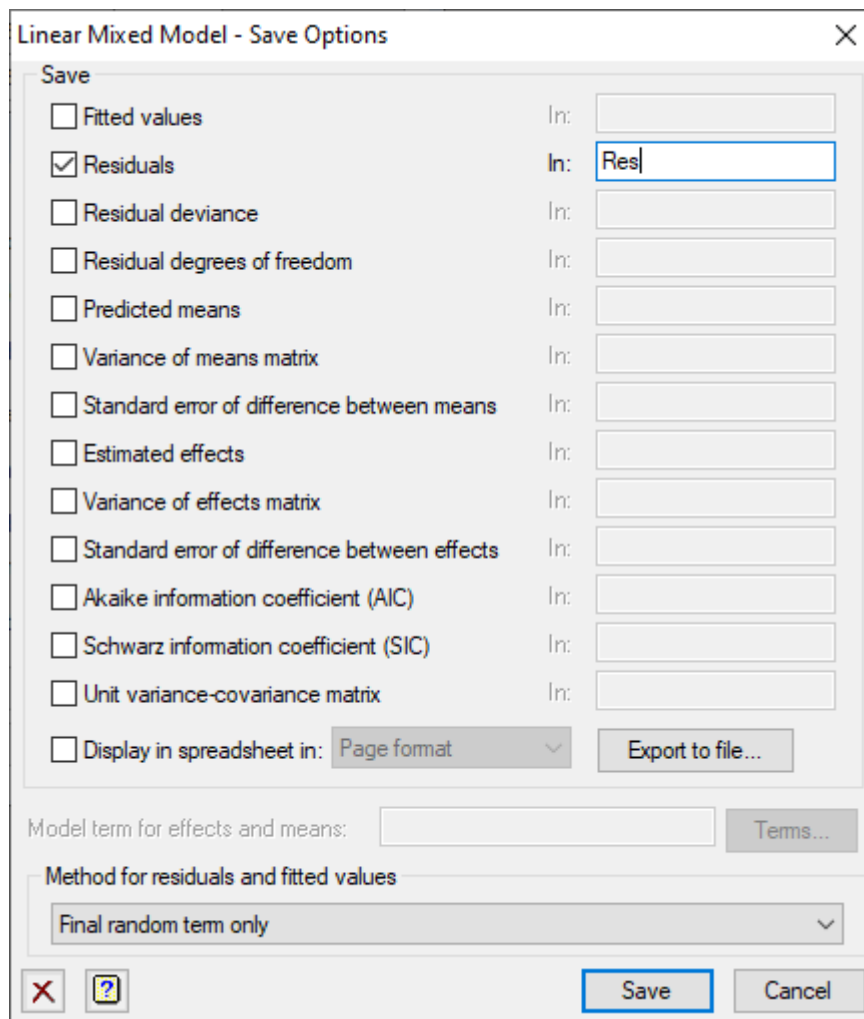
Figure 3.17: Saving residuals via Save button on Spatial Model - Regular Grid menu.

It can be very helpful to plot the residuals against the rows and columns of the layout in order to detect spatial patterns. We can do this using a trellis plot (Graphics | Trellis Plot) defined as shown in Figure 3.18. We have specified that the residuals (Y-values: `Res`) be plotted against the row numbers (X-values: `Row`), forming a separate plot for each column (Groups for frames: `Column`). We have asked for the plotting to be done using both points and lines, so that the points will be interpolated by straight lines. We have also added axis titles and set the label spacing using the X Axis and Y Axis tabs (not shown). The resulting plot is shown in Figure 3.19.
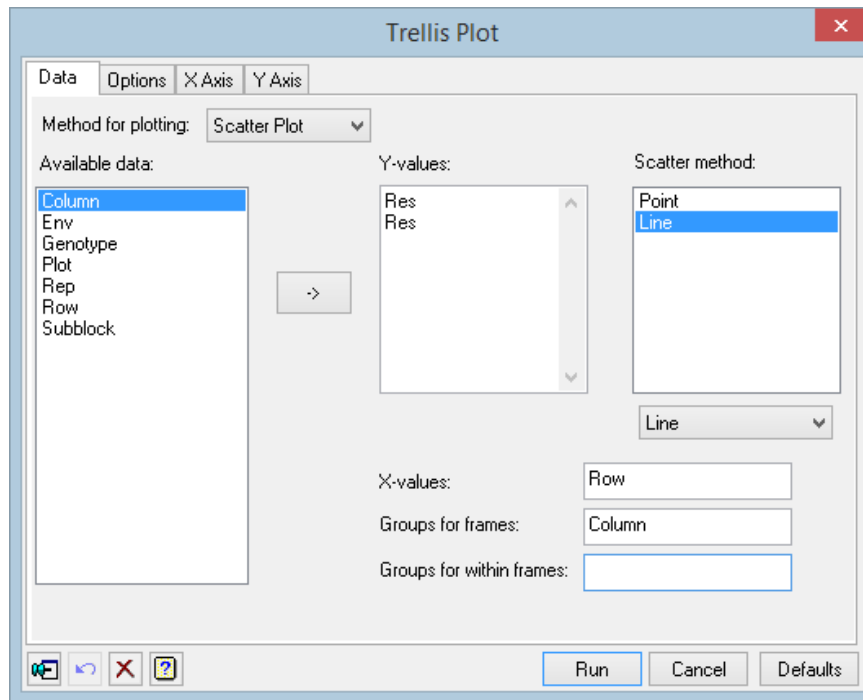
Figure 3.18: Specifying a trellis plot of residuals.
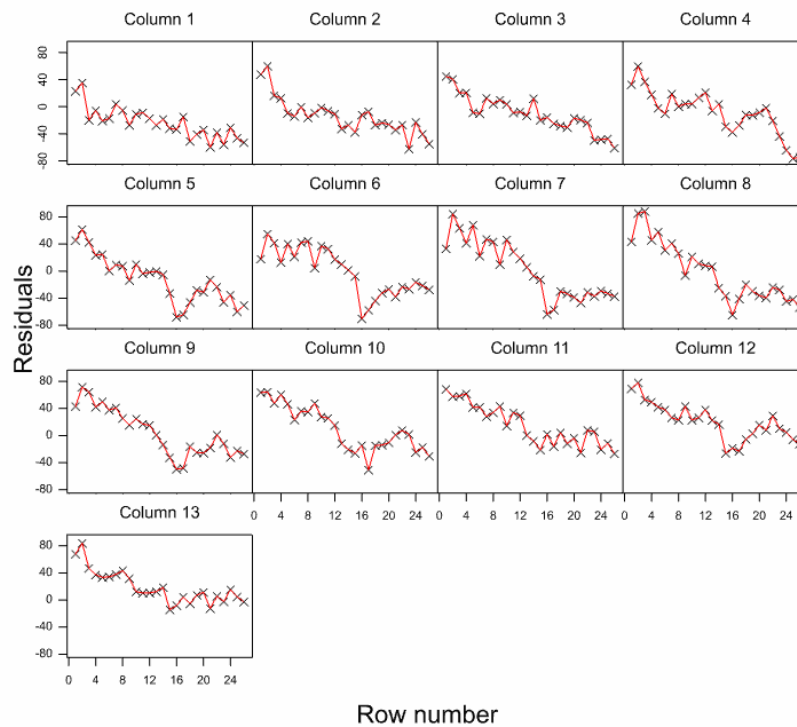
Figure 3.19: Plot of residuals from exploratory model against row number, for each column.

There is a clear linear trend across rows (Figure 3.19). This trend is reflected in the large variance component obtained for factor `Rep` in the design-based analysis of Section 3.1. However, the factor `Rep` fits a single common effect for the whole of the first replicate, then another effect for the whole of the second replicate, and this does not account for the linear trend. The term `Rep.Subblock` does fit a separate effect for each individual row in the design-based model, and this term will account for the remaining trend, as shown in Figure 3.20. But the assumption behind the `Rep.Subblock` effects is that they are independent and uncorrelated, which is clearly not the case here. This is an example of so-called global trend, a strong trend that extends across the whole trial, and this is best accounted for by modelling the linear trend directly. This term can be added to the model by checking the Linear trend across rows box on the Spatial Model - Regular Grid menu (Figure 3.16).
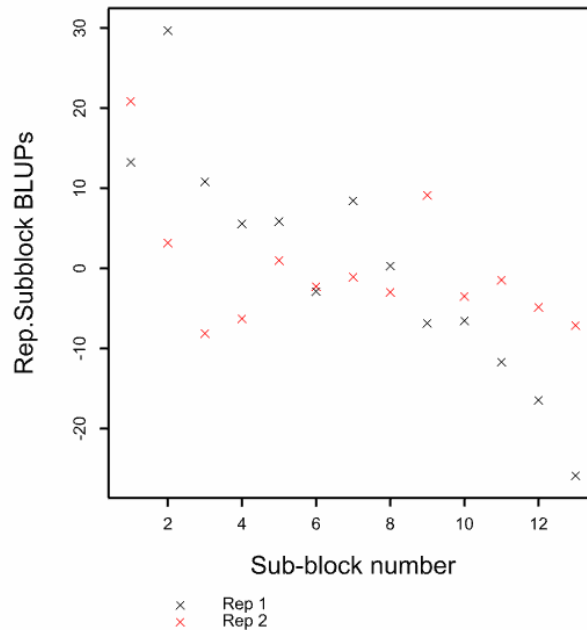
Figure 3.20: BLUPs for `Rep.Subblock` term from design-based analysis of the `HEAT05` trial (Section 3.1).

Likewise, a trellis plot can be used to plot the residuals from the exploratory model against column number, separately for each row (Figure 3.21). This plot also shows a trend running across columns, and we can model this global trend using a linear term by checking Linear trend across columns.
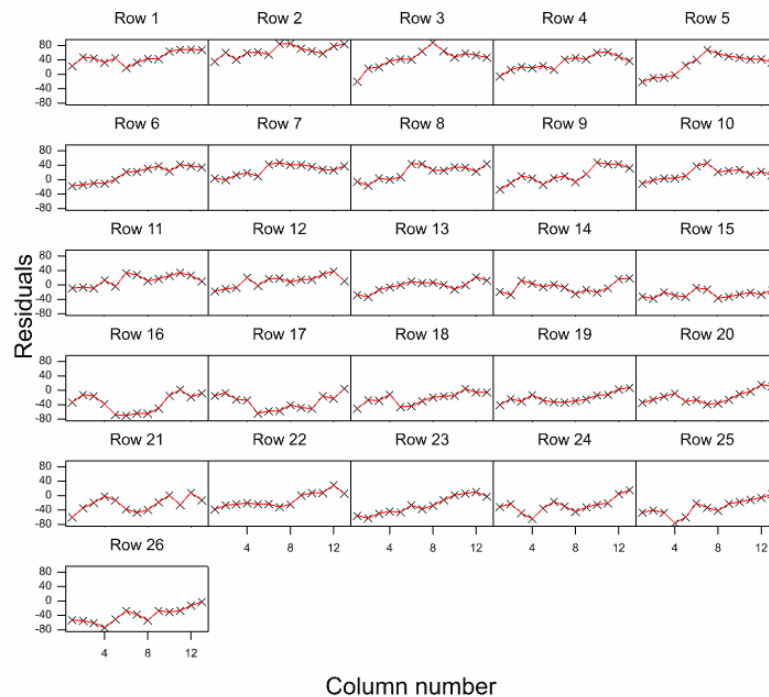
Figure 3.21: Plot of residuals from exploratory model against column number, for each row.

Before fitting a new model, we will confirm these conclusions using additional residual plots obtained directly from the Spatial Model - Regular Grid | Further Output menu (Figure 3.22). From the Residual Plots menu, we could obtain shade (or contour) plots of the residuals in field layout, as done in Section 3.1 & Figure 3.8. From the Display Variogram menu, we can generate a two-dimensional variogram (Figure 3.22), as described by Stefanova *et al.* (2009).

A variogram is a diagnostic tool used to investigate patterns of spatial correlation in plot residuals, commonly used in geo-statistics and applied to two-dimensional trends in field trials by Gilmour *et al.* (1997). The value of the two-dimensional variogram at position (i,j) is the average squared difference for all pairs of residuals from plots that are i rows and j columns apart. The variogram value at the origin (0, 0) is always zero. Stefanova *et al.* (2009) give examples of typical patterns seen in practice. A global trend across the trial generates a trend across the range of the variogram. Row (or column) effects tend to generate systematic patterns in the row (or column) direction. Serial correlation across rows tends to generate a smooth curve in the variogram in the row

direction, increasing away from the origin up to a plateau. Uncorrelated residuals will generate a rough but flat plateau, apart from the zero value at the origin.
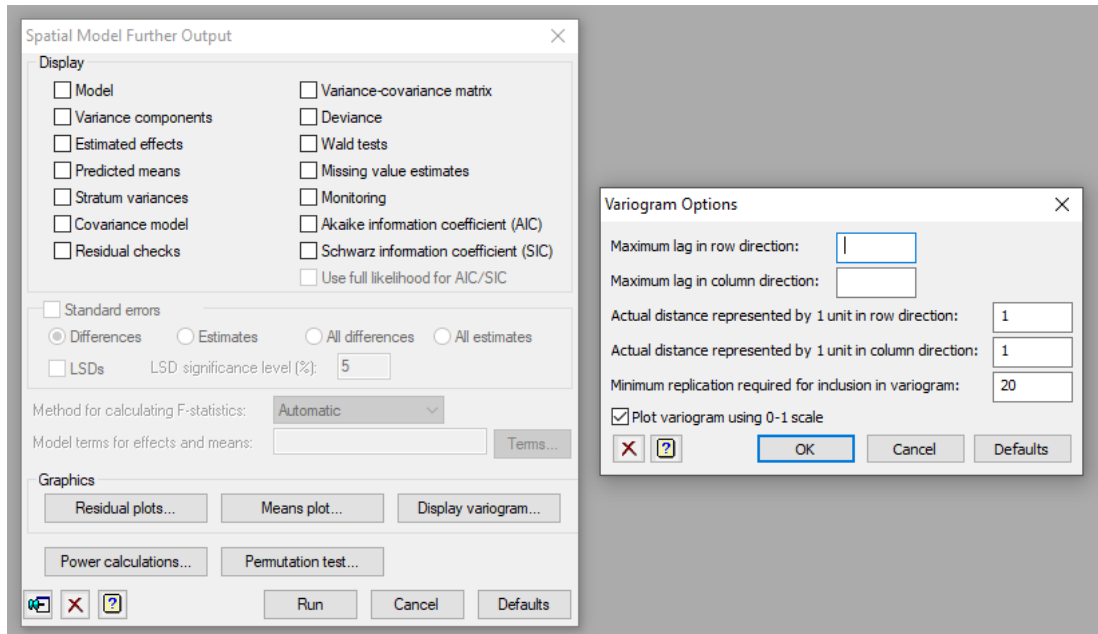
Figure 3.22: Spatial Model Further Output menu with Variogram Options.

The variogram generated from our exploratory model is shown in the left-hand-side of Figure 3.23. This variogram does not plateau, reflecting the presence of linear trend in both the row and column directions.
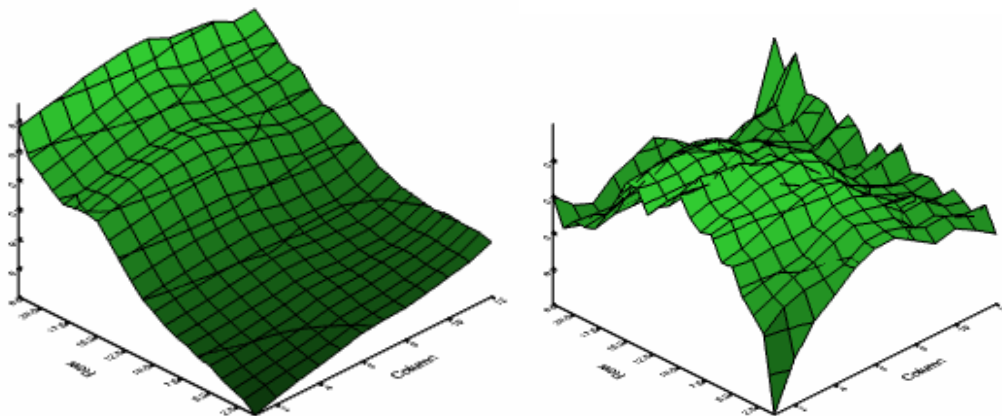
Figure 3.23: Sample variogram from exploratory model (left) and with linear row and column effects added (right).

If we re-fit the model with linear row and column effects added, we obtain the variogram on the right-hand-side of Figure 3.23: this variogram does now appear to plateau (apart from a little noise) and the smooth increase suggests the presence of serial correlation in both directions. Residual plots from this model (not shown) indicate no evidence of any remaining global trends. There is no suggestion of the presence of row and column effects either from the variogram (no systemetic patterns in either direction) or residual plots. A summary of the model, estimated variance parameters and Wald tests from this model (Options | Display then check Model and Variance components from the Spatial Model - Regular Grid menu) are shown in the output below:

```
REML variance components analysis
=================================

Response variate:  yield
Fixed model:       Constant + lin_row + lin_col
Random model:      Row.Column + Genotype
Number of units:   338

Row.Column used as residual term with covariance structure as below

Sparse algorithm with AI optimisation
All covariates centred


Covariance structures defined for random model
----------------------------------------------

Covariance structures defined within terms:

Term                    Factor    Model                        Order  No. rows
Row.Column              Row       Auto-regressive (+ scalar)       1        26
                        Column    Auto-regressive                  1        13


Estimated variance components
-----------------------------

Random term             component        s.e.
Genotype                   1391.7        166.4


Residual variance model
-----------------------
```

```
Term            Factor      Model(order)  Parameter        Estimate      s.e.
Row.Column                                Sigma2             336.1      43.1
                Row         AR(1)         phi_1             0.2894    0.0835
                Column      AR(1)         phi_1             0.4939    0.0700



Tests for fixed effects
-----------------------


Sequentially adding terms to fixed model

Fixed term              Wald statistic  n.d.f.   F statistic  d.d.f.    F pr
lin_row                         153.60       1        153.60    27.2  <0.001
lin_col                          42.76       1         42.76    35.7  <0.001

Dropping individual terms from full fixed model

Fixed term              Wald statistic  n.d.f.   F statistic  d.d.f.    F pr
lin_row                         149.87       1        149.87    27.2  <0.001
lin_col                          42.76       1         42.76    35.7  <0.001

* MESSAGE: denominator degrees of freedom for approximate F-tests are calculated
using   algebraic   derivatives   ignoring   fixed/boundary/singular   variance
parameters.
```

From the model summary, we can see that the menu has generated term `Row.Column`, used as the residual term, to which the spatial model is applied. The Wald tests for the linear trend across both rows (`lin_row`) and columns (`lin_col`) are highly significant, reflecting the strong trend observed (background information on Wald tests is given in Section 10.6). The correlation parameters for the AR1 processes are 0.289 for rows and 0.494 for columns, suggesting the presence of some local spatial trend.

We are now ready to build a model for this data. We will re-introduce the design structure as random terms `Rep/Subblock` in addition to the spatial model at the plot level and the `Genotype` factor. Linear row and column trends will still be fitted as fixed. This model specification is shown in Figure 3.24, and we have chosen to display the model summary, variance components, deviance and information criteria to obtain the output below:
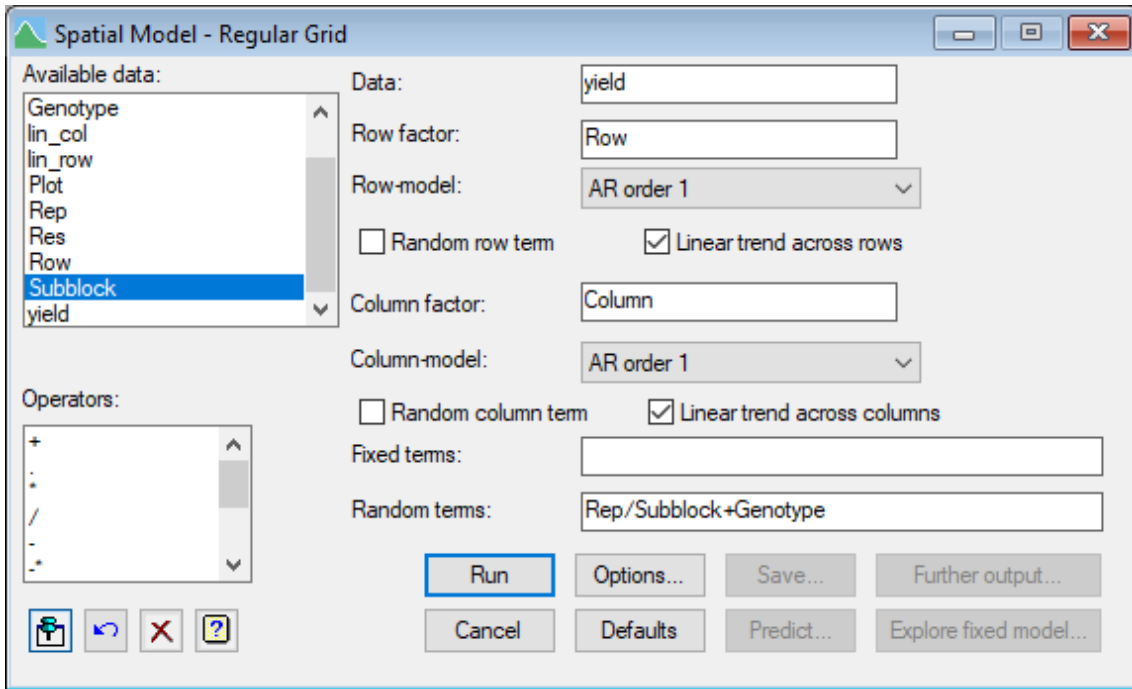
Figure 3.24: Settings for spatial model for `HEAT05` trial including design factors.

```
REML variance components analysis
=================================


Response variate:   yield
Fixed model:        Constant + lin_row + lin_col
Random model:       Row.Column + Rep + Rep.Subblock + Genotype
Number of units:    338


Row.Column used as residual term with covariance structure as below


Sparse algorithm with AI optimisation
All covariates centred


Covariance structures defined for random model
----------------------------------------------


Covariance structures defined within terms:

Term                 Factor      Model                          Order   No. rows
Row.Column           Row         Auto-regressive (+ scalar)         1         26
                     Column      Auto-regressive                    1         13
```

```
Estimated variance components
-----------------------------

Random term              component        s.e.
Rep                          22.4         82.3
Rep.Subblock                 23.2         29.8
Genotype                   1392.8        166.5


Residual variance model
-----------------------

Term           Factor      Model(order)  Parameter        Estimate      s.e.
Row.Column                               Sigma2            304.6       47.2
               Row         AR(1)         phi_1           0.2665     0.0881
               Column      AR(1)         phi_1           0.4425     0.0944



Deviance: -2*Log-Likelihood
---------------------------

                Deviance   d.f.
                 2638.51   329

Note: deviance omits constants which depend on fixed model fitted.


     Akaike information coefficient     2650.51
Schwarz Bayes information coefficient   2673.40

Note: omits constants, (n-p)log(2pi) - log(det(X'X)), that depend only on the
fixed model.

(based on the residual log-likelihood)
```

The inclusion of the `Rep/Subblock` design structure has had little impact on the variance parameters. The deviance and information criteria can be used to compare random models (for details see Section 10.4). We would not drop the design structure, but we might investigate whether the small correlation across rows is really improving the variance model. We can drop this correlation from the model by setting Row-model: Identity on the Spatial Model - Regular Grid menu. We re-fit and obtain the following values of the deviance and information criteria:

```
Deviance: -2*Log-Likelihood
---------------------------
```

```
              Deviance    d.f.
               2644.17    330

Note: deviance omits constants which depend on fixed model fitted.


        Akaike information coefficient      2654.17
Schwarz Bayes information coefficient      2673.25

Note: omits constants, (n-p)log(2pi) - log(det(X'X)), that depend only on the
fixed model.

(based on the residual log-likelihood)
```

This model is nested within the previous model, by setting the row correlation parameter equal to zero, so we can use a likelihood ratio test to compare the two models. The deviance has increased by 5.66 units, which is large compared to a chi-square distribution on 1 df (95th percentile = 3.84) and so we conclude that this parameter, although small, has significantly improved the fit of the model. The AIC has also increased, although the SIC has decreased very slightly. For nested models, we would usually use the likelihood ratio test based on the change in deviance in preference to the information criteria, and so we do not change our conclusions. The increase in deviance is even larger (as we might expect) if we set the column correlation parameter equal to zero. We therefore conclude that the previous model, including the design structure as well as separable spatial trend over both rows and columns, with global linear trend for rows and columns, gives an adequate model for this trial. We can then go on to use this model to produce predicted means for QTL analysis.

3.7   Extension to multi-trait data sets

The PSEA menu can also be used for preliminary analysis of several traits for the same trial, prior to a multi-trait analysis (see Chapter 7). In this case, the predicted trait means should be kept as separate data sets, and not combined together. However, these separate preliminary analyses ignore correlations across measurements from the same plots, whereas better predictions may sometimes be obtained from a multivariate analysis, particularly when some data is missing. An alternative approach is a joint analysis of the traits using the Stats | Mixed Models (REML) | Multivariate Linear Models menu, followed by prediction (Predict), then the predicted trait means can be imported directly into the QTL Data Space.

3.8　References

Cullis, B.R., Smith, A.B., & Coombes, N.E. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological & Environmental Statistics*, **11**, 381-393.

Gilmour, A.R., Cullis, B.R., & Verbyla, A.P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural Biological and Environmental Statistics*, **2**, 269-293

Kearsey, M.J., & Pooni, H.S. (1996). The genetical analysis of quantitative traits. Chapman and Hall, London.

Möhring, J., & Piepho, H.-P. (2009). Comparison of weighting in two-stage analysis of plant breeding trials. *Crop Science*, **49**, 1977-1988.

Smith, A., Cullis, B., & Gilmour, A. (2001). The analysis of crop variety evaluation data in Australia. *Australian and New Zealand Journal of Statistics*, **43**, 129-145.

Stefanova, K.T., Smith, A.B., & Cullis, B.R. (2009). Enhanced diagnostics for the spatial analysis of field trials. *Journal of Agricultural, Biological and Environmental Statistics*, **14**, 392-410.

Welham, S.J., Gogel, B.J., Smith, A.B., Thompson, R., & Cullis, B.R. (2010). A comparison of analysis methods for late-stage variety evaluation trials. *Australian & New Zealand Journal of Statistics*, **52**,125-149.

4 Multi-environment trial analyses: modelling genotype by environment interaction

A multi-environment trial (MET) comprises a series of experiments conducted over a variety of environmental conditions. Plant breeding programmes generate MET data when evaluating genotypes (i.e. cultivars) across a range of geographic locations, and possibly over a number of years (or seasons). The set of trials within and across years is designed to provide a range of growing conditions, and the term "*environments*" is used to describe these conditions (i.e. the combinations of locations and years). Of interest is to identify genotypes that perform well in all environments (i.e. broad adaptation). However, the relative performance of the genotypes can change between environments; known as *genotype by environment* (G×E) *interaction*. The G×E phenomenon results from different genotypes responding to environmental variation in different ways. In QTL analysis, a key objective is to determine if QTL effects are consistent across environments, or whether there are environmental interactions (QTL×E).

The identification of QTLs from MET data depends on the appropriate modelling of the G×E interaction. In this chapter, we illustrate the use of a mixed model to quantify and describe the G×E interaction. We outline the underlying statistical theory for modelling MET data, including the different variance-covariance models available for the G×E matrix in Genstat (Section 4.1). Then, we demonstrate the use of the Select Best Variance-covariance Model menu to select the best variance-covariance model for subsequent use in linkage analysis (Section 4.2), and describe how to fit weights to accommodate within-trial plot variation (Section 4.3). Finally, we illustrate the use of other exploratory tools (AMMI and GGE biplots) to investigate the structure of the G×E interaction (Section 4.4).

In this chapter you will learn how to perform a G×E analysis for a multi-environment trial data set, including:

- the different variance-covariance models available for the G×E matrix in Genstat (Section 4.1)
- how to select the best variance-covariance model for use in linkage analysis (Section 4.2)
- how to accommodate the within-trial plot variation (Section 4.3)
- how to use other exploratory tools for analysis of G×E variation: AMMI and GGE biplots (Section 4.4).

4.1   Modelling genotype by environment interaction

A statistical analysis of MET data aims to provide reliable predictions of genotype performance across environments. In this section, we describe the use of a linear mixed model to quantify and characterize the G×E interaction. We begin by illustrating the G×E phenomenon (Section 4.1.1; also see Malosetti *et al*., 2013) before discussing the underlying linear mixed model framework for modelling a MET (Section 4.1.2; also see Smith *et al.*, 2005), including the different variance-covariance models available in Genstat to describe the variation between genotypes both within and across environments (Section 4.1.3).

In Genstat, G×E analysis is performed on the table of G×E trait means (with their unit errors if available). This is the so-called *two-stage strategy* for analysing MET data (see Section 4.2).

4.1.1   Genotype by environment interaction

G×E interaction occurs when the relative phenotypic performance of a set of genotypes depends on the environment. Whereas some genotypes may perform well across a wide range of environmental conditions (i.e. broadly adapted genotypes), others perform well in only a subset of environments (i.e. specifically adapted genotypes).

To illustrate G×E interaction, consider the phenotypic response of two genotypes across two environments. The four possible patterns of phenotypic response are shown in Figure 4.1. The first pattern (Figure 4.1a) is of no interaction, where the effects of genotype and environment are independent of one another (i.e. behave additivity). That is, the difference in phenotypic response between the two genotypes is the same across the two environments. In this case, `Genotype 2` yields more than `Genotype 1` by a constant amount in both environments. The three remaining patterns (Figure 4.1b-d) are non-additive. Here the difference between genotypes changes between environments, i.e. genotypic performance is dependent on environment. The cross-over interaction (Figure 4.1d) is the most critical for breeders. It implies that the choice of the best genotype is determined by the environment, seriously hampering efforts to select genotypes that perform well across a range of environmental conditions. When the ordering of environments has a biological interpretation, the divergence pattern in Figure 4.1b is perhaps more common.
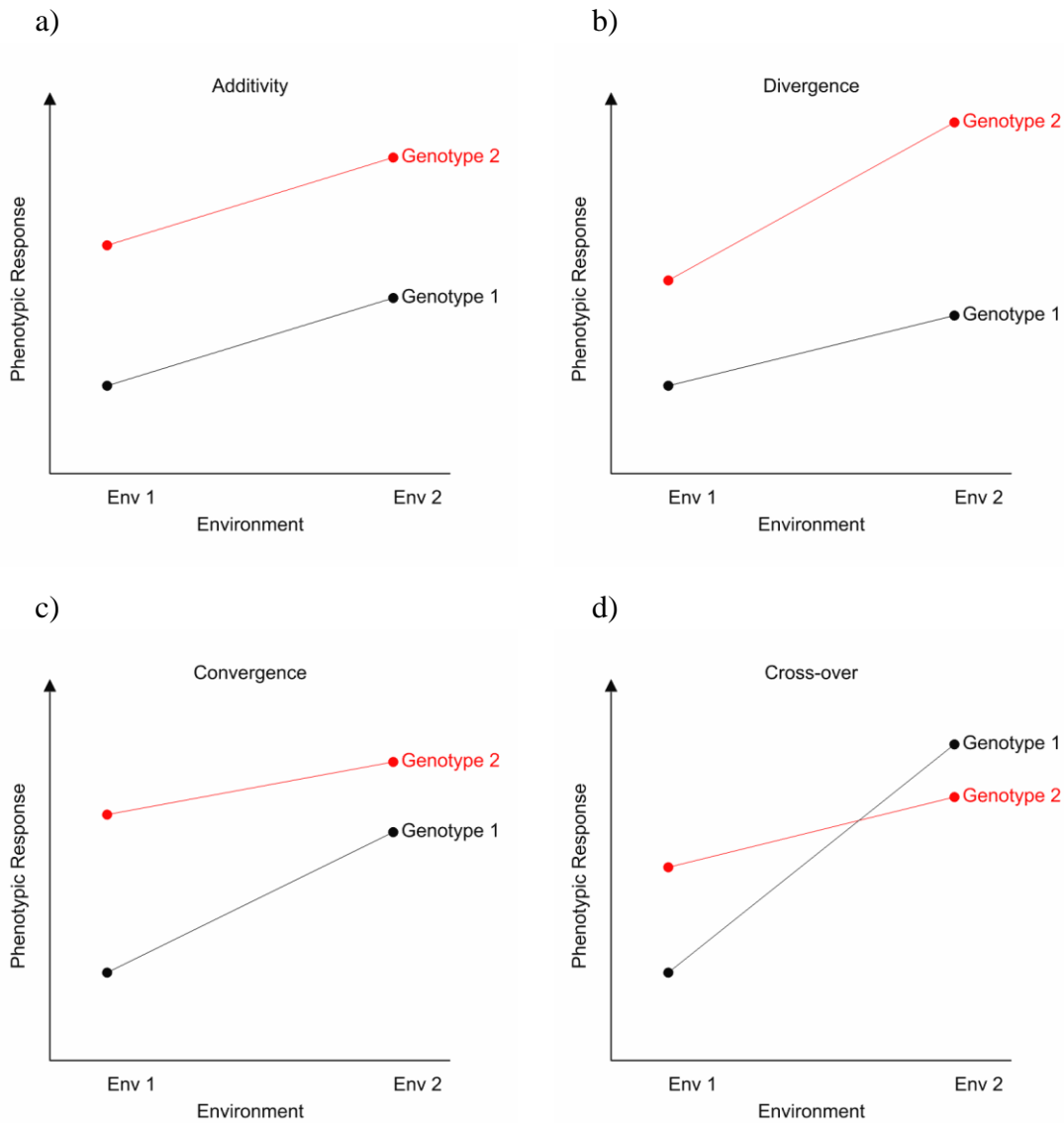
Figure 4.1: Patterns of G×E interaction: a) additivity, b) divergence, c) convergence, and d) cross-over.

Using Figure 4.1, we have considered the effect of G×E interaction on relative changes in mean phenotypic response. However, G×E interaction can also have consequences on genetic variance and correlation (or covariance). For example, when the G×E interaction is large, the phenotypic performance of a set of genotypes in one environment may not be very informative about their performance in another, very different environment. This results in a low genetic correlation. Only those environments with similar characteristics lead to a strong genetic correlation. Furthermore, G×E interaction can induce heterogeneity of genetic variance across environments, where the magnitude of genetic

variance within an individual environment is different between environments (Malosetti *et al.,* 2013).

In Figure 4.2 we illustrate the effects of G×E interaction on genetic variance and correlation between two environments. In the case of Figure 4.2a, the variation between genotypes is similar in both environments (i.e. homogeneous). However, as there is no consistent pattern in relative genotypic performance between the two environments, the genetic correlation is low. In Figure 4.2b, the correlation between the two environments is negative, as the cross-overs mean that genotypes high in `Env 1` will be low in `Env 2`, but it is clear that the variation between genotypes in `Env 1` is substantially smaller than `Env 2` - this is variance heterogeneity.

a)   b)
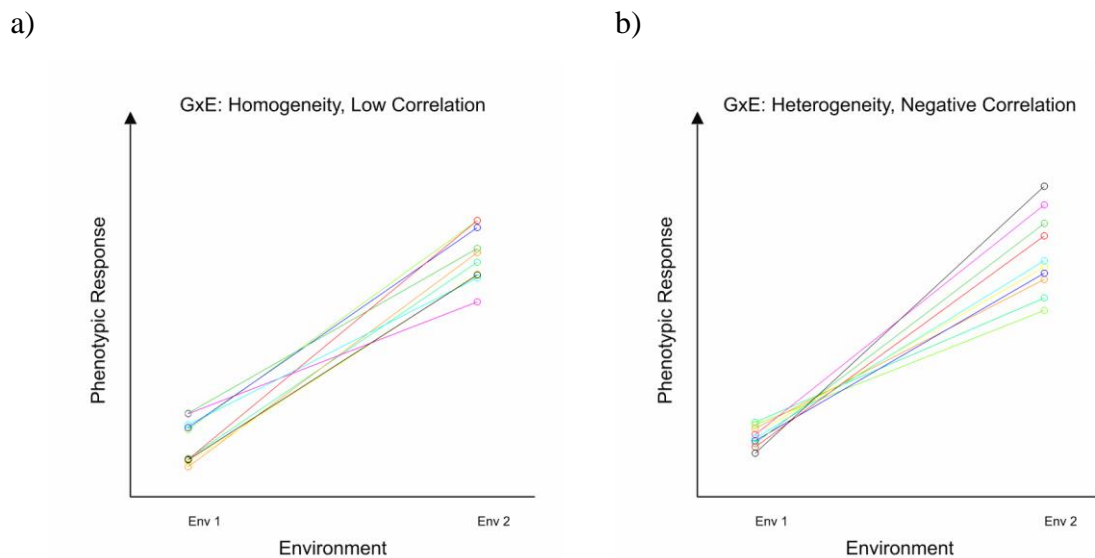
Figure 4.2: Examples illustrating the effects of G×E interaction on genetic variance and correlation.

4.1.2   General model for the analysis of MET data

There are a number of possible models for MET data. These models are built by stacking the data vectors for each individual trial (i.e. environment). The basic model consists of genotype and environment effects plus their interaction. In their review paper, Smith *et al*. (2005) discuss the classification of genotype effects as fixed or random. They follow an approach motivated by a quantitative genetics interpretation of the G×E interaction. As explained by Falconer & Mackay (1996), the phenotypic response in each environment can be regarded as a different "*character*". If the genetic correlation between environments is high, then phenotypic performance in the different environments

represents essentially the same character. Conversely, if it is low, then the characters are very different. In other words, genetic correlation between phenotypic responses in different environments provides information on the similarity of these characters. This leads naturally to classifying the genotype and G×E interaction effects as random, with independence between genotypes and some variance-covariance matrix across environments. Environments are then fitted as fixed effects in the linear mixed model. This is the approach taken by Genstat.

In Genstat, the analysis of MET data is performed on the trait means for each genotype; known as a two-stage analysis (see Section 4.2). These are obtained from the separate analyses of the individual trials (as described in Chapter 3). The model for the observed entry, y_{ij}, in the G×E table of genotype i ($i=1, …, n$) by environment j ($j=1, …, m$) trait means is:

$$y_{ij} = E_j + u_{ij} + e_{ij}$$ Equation 1

where E_j represents the fixed effect of environment j, u_{ij} the random effect of genotype i in environment j (i.e. the combined genotype and interaction effect), and e_{ij} the error in the data arising from within-trial variation, known as the "*unit error*". If estimates of unit error are not available, this term is omitted from the model.

The G×E random effects, u_{ij}, are assumed to be Normally distributed with mean 0 and variance-covariance structure VCOV($\boldsymbol{u}_{ij}$). An important step in G×E analysis is selecting an appropriate variance-covariance model to portray the variation between genotypes both across and within environments. The selected model is used as the default variance-covariance model in subsequent QTL analysis. In the following section, we describe the variance-covariance models available in Genstat.

4.1.3 Variance-covariance models

A statistical analysis of MET data aims to provide reliable predictions of genotype performance across environments. This requires selecting an appropriate variance-covariance model to describe the variation between genotypes both across and within environments. For example, Figure 4.2b illustrates the need to allow for heterogeneity of variance between environments.

To help you decide which variance-covariance models best suits your data set and analysis needs we described them in turn. For each model, the structure of the variance-covariance matrix, VCOV($\boldsymbol{u}_{ij}$), is summarized in Table 4-1.

4.1.3.1   Identity

The simplest model, the identity model, assumes that the within-environment variances are all the same, and that the environment covariances (or correlations) are equal to 0. That is, this model does not allow for heterogeneity of genetic variance across environments nor genetic correlations between environments. In practice, this model is rarely realistic.

4.1.3.2   Compound symmetry

The compound symmetry model (also known as the uniform model) partitions the effect u_{ij} into two parts, $u_{ij} = G_i + (GE)_{ij}$, one corresponding to the random genotype main effect and the other to the residual (which includes the true G×E interaction and the residual error). G_i and $(GE)_{ij}$ are fitted as random terms with independent effects and variance components var$(G_i)=\sigma_g^2$ and var$((GE)_{ij})= \sigma_{ge}^2$, respectively. It generates a uniform covariance structure, with equal variances and covariances across environments. That is, the variation between genotype effects is the same within each environment, and the covariance across environments is the same for every pair of environments.

   The compound symmetry model has traditionally been used to model MET data from plant breeding trials. However, in practice it is unrealistic as genotype variation tends to differ among environments, giving heterogeneity of genetic variance (Figure 4.2b), and some environments are more alike than others, resulting in unequal correlations between pairs of environments.

4.1.3.3   Diagonal

The diagonal model allows a separate variance for each environment, $(\sigma_{ge_j}^2 : j = 1, ..., m)$ but the covariances between environments are set to 0. This accommodates heterogeneity of genetic variance across environments but does not model the genetic covariances (or correlations) between environments.

4.1.3.4   Uniform covariance with unequal variances

The uniform covariance with unequal variance model (sometimes also called the heterogeneous compound symmetry model) is a combination of the compound symmetry and diagonal models. G_i is fitted as a random term and the individual environment variances $(\sigma_{ge_j}^2 : j = 1, ..., m)$ are modelled, allowing for heterogeneity of genetic variance across environments. However, a common covariance between environments is assumed.

4.1.3.5   Uniform correlation with unequal variances

The uniform correlation with unequal variance model is similar to uniform covariance with unequal variance model but assumes a common correlation between environments. The heterogeneity of individual environment variances is accommodated.

4.1.3.6   Factor analytic of order k

The factor analytic models are multiplicative models that parsimoniously accommodate heterogeneity of genetic variance across environments and model the genetic covariances between environments. This model therefore captures the nature of heterogeneous variances and covariances found to occur in most MET data. A factor analytic model of order k uses k factors to describe the variance-covariance matrix, VCOV($\boldsymbol{u}_{ij}$). See Smith *et al.*, 2001b for details. The QTL menus allow for $k = 1$ or 2. Note, `FA2` is not available for less than 5 environments.

4.1.3.7   Unstructured

The most general form of the variance-covariance matrix, VCOV($\boldsymbol{u}_{ij}$), which does not impose constraints on correlations in genetic performance across environments, is of a fully unstructured form. However, it is not parsimonious and may be difficult to fit when the number environments (m) is large or the number of genotypes (n) is small.

Table 4-1: Variance-covariance models for VCOV($\boldsymbol{u}_{ij}$).

Description	Abbreviation in output	Variance-covariance matrix	Variance	Covariance	Number of parameters
Identity	`identity`	$\sigma_{ge}^2\mathbf{I}$	homogeneous σ_{ge}^2	none 0	1
Compound symmetry	`cs`	$\sigma_g^2\,\mathbf{J} + \sigma_{ge}^2\,\mathbf{I}$	homogeneous $\sigma_g^2 + \sigma_{ge}^2$	homogeneous σ_g^2	2
Diagonal	`diagonal`	$\mathbf{D}$	heterogeneous $\sigma_{ge_j}^2$	none 0	m
Uniform covariance, unequal variances	`hcs`	$\sigma_g^2\,\mathbf{J} + \mathbf{D}$	heterogeneous $\sigma_g^2 + \sigma_{ge_j}^2$	homogeneous σ_g^2	$m + 1$
Uniform correlation, unequal variances	`outside`	$\sqrt{\mathbf{D}}\,\mathbf{K}\,\sqrt{\mathbf{D}}$	heterogeneous $\sigma_{ge_j}^2$	homogeneous σ_g^2	$m + 1$
Factor analytic order *1*	`fa`	$\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \mathbf{D}$	heterogeneous $\lambda_{1j}^2 + \sigma_{ge_j}^2$	heterogeneous $\lambda_{1j}\lambda_{1j*}$	$2m$
Factor analytic order *2*	`fa2`	$\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \mathbf{D}$	heterogeneous $\lambda_{1j}^2 + \lambda_{2j}^2 + \sigma_{ge_j}^2$	heterogeneous $\lambda_{1j}\lambda_{1j*} + \lambda_{2j}\lambda_{2j*}$	$3m$
Unstructured	`unstructured`	$\sqrt{\mathbf{D}}\,\mathbf{K}\,\sqrt{\mathbf{D}}$	heterogeneous $\sigma_{ge_j}^2$	heterogeneous σ_{jj*}^2	$m(m + 1)/2$

m is the number of environments.

σ_g^2 and σ_{ge}^2 are variance components for the genotype and G×E interactions random effects, respectively.

I is an identity matrix of size *m*.

J is an $m \times m$ matrix of ones.

D is a diagonal matrix containing environment specific variances ($\sigma_{ge_j}^2$: *j = 1, ..., m*).

K is an $m \times m$ matrix of ones on the diagonal and θ_{jj*} on the off-diagonals, where θ_{jj*} is the correlation between environment *j* and *j**.

$\boldsymbol{\Lambda}$ is an $m \times k$ matrix of loadings, λ_{kj}, from a factor analytic model of order *k*.

4.2   Genotype-by-environment analysis

In Genstat, G×E analysis of MET data is performed in two-stages (see Smith *et al.*, 2001a; Welham *et al.*, 2010). In the first stage (stage I), trait means (y_{ij}) and unit errors (e_{ij}) for each genotype are obtained from the separate analysis of the individual trials comprising the MET (refer to Chapter 3). These are then combined in an overall mixed model analysis in the second stage (stage II). In this section, we illustrate stage II of the G×E modelling process using trait means from the 8 environment CIMMYT maize trials (Section 1.3.2) held in file `F2maize_pheno.csv`.

The stage II analysis may be unweighted (e.g. Patterson & Silvey, 1980) or weighted, using the unit error, to reflect the relative precision of genotype means from each trial (e.g. Smith *et al.,* 2001a). Weighted analyses are discussed in Section 4.3.

If raw plot (unit) data are available refer to Chapter 3, Section 3.2, to obtain the trait means and unit errors. Before embarking on G×E analysis, exploratory data analysis is recommended (see Section 2.4.1).

G×E analysis is performed using the Select Best Variance-covariance Model menu (Figure 4.3) accessed via Stats | QTLs (Linkage/Association) | Phenotypic Analysis | Select Best Variance-covariance model for Multiple Environments; or, from the QTL Data View shortcut Phenotypic analysis | Select Best Variance-covariance model for Multiple Environments.
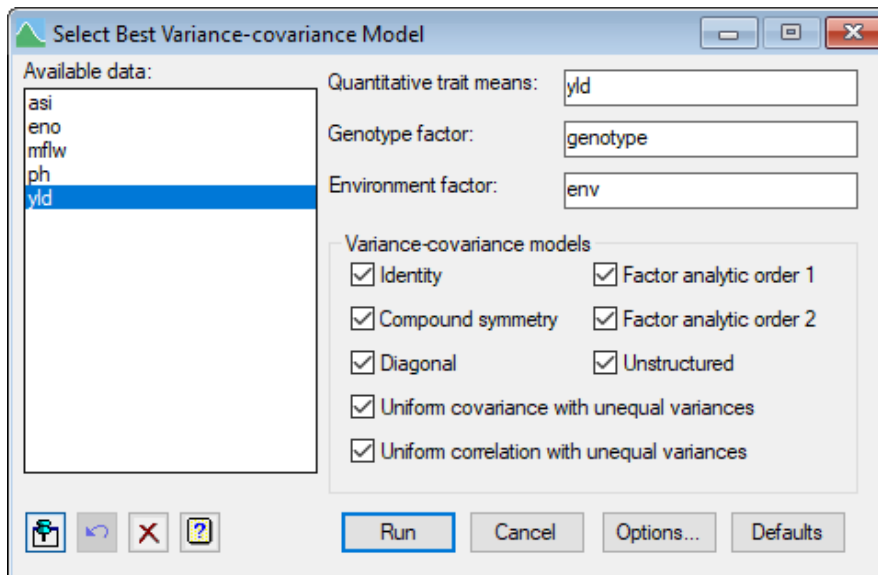
Figure 4.3: Select Best Variance-covariance Model window for analysis of `yld` trait means from the `F2maize_pheno.csv` data set.

The fields for Quantitative trait means:, Genotype factor: and Environment factor: are automatically filled using available data in the QTL Data Space. We set the Quantitative trait means: field to `yld`.

Different models can be fitted that assume different variance-covariance structures (see Section 4.1.3), and their goodness of fit compared to select the best model. The appropriate set of variance-covariance models to compare will depend on the complexity of the data set and the aims of your analysis. For the purpose of illustration, we compare all available models (Figure 4.3).

Clicking on the Options button opens the Select Variance-covariance Model Options window (Figure 4.4) where you can specify what output to display, which information criterion to use for comparing the different models, and whether to include unit errors in the analysis (see Section 4.3). The default is to provide the Summary table of the variance-covariance models fitted and use the Schwarz information criterion (SIC) to select the best model. We also request output from the best model to be displayed by selecting Best model.
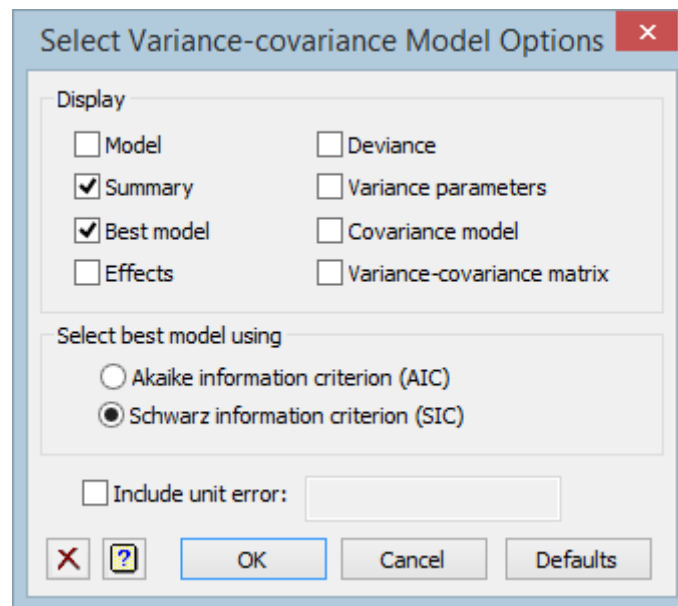
Figure 4.4: Options for the Select Best Variance-covariance Model menu.

The output from the analysis of `yld` is given below:

```
Summary
=======


Trait: yld

          Model        AIC        SIC    Deviance NParameters
             FA      17471      17524      17439          16
            FA2      17455      17532      17409          23
        OUTSIDE      17523      17554      17505           9
   UNSTRUCTURED      17456      17577      17384          36
            HCS      17692      17722      17674           9
             CS      17918      17924      17914           2
       DIAGONAL      17906      17933      17890           8
       IDENTITY      18287      18290      18285           1


Best model: FA (on basis of criterion SIC)




Residual variance model
-----------------------


Term            Factor        Model(order)  Parameter        Estimate        s.e.
genotype.env                                Sigma2             1.000       fixed
                genotype      Identity      -                     -           -
                env           FA(1) (covariance form)
                                            g_11              75.02       10.59
                                            g_21              108.8         9.3
                                            g_31              115.5         9.2
                                            g_41              28.53        4.71
                                            g_51              19.37        4.60
                                            g_61              122.6        13.3
                                            g_71              83.61        8.72
                                            g_81              108.0         9.7
                                            psi_1            16910.       1752.
                                            psi_2             9544.       1191.
                                            psi_3             8527.       1146.
                                            psi_4             3540.        360.
                                            psi_5             3535.        351.
                                            psi_6            24025.       2633.
                                            psi_7            10176.       1127.
                                            psi_8            11013.       1321.




Estimated covariance models
---------------------------


Variance of data estimated in form:

V(y) = Sigma2.R
```

```
where: V(y) is variance matrix of data
       Sigma2 is the residual variance
       R is the residual covariance matrix


Residual term: genotype.env

Sigma2: 1.000

R uses direct product construction

Factor: genotype
Model:  Identity ( 211 rows)

Factor: env
Model:  FA (covariance)

Covariance matrix:


1   22537
2    8159   21373
3    8662   12558   21860
4    2140    3103    3295    4354
5    1453    2107    2237     553    3910
6    9197   13335   14157    3498    2375   39057
7    6272    9093    9654    2386    1620   10251   17166
8    8105   11751   12475    3083    2093   13246    9033   22686
        1       2       3       4       5       6       7       8



Correlation matrix:
        HN96b       1.0000
        IS92a       0.3717      1.0000
        IS94a       0.3902      0.5810      1.0000
        LN96a       0.2161      0.3217      0.3377      1.0000
        LN96b       0.1548      0.2305      0.2419      0.1340      1.0000
        NS92a       0.3100      0.4615      0.4845      0.2683      0.1922
        SS92a       0.3189      0.4747      0.4983      0.2759      0.1977
        SS94a       0.3584      0.5336      0.5602      0.3102      0.2222
                    HN96b       IS92a       IS94a       LN96a       LN96b

        NS92a       1.0000
        SS92a       0.3959      1.0000
        SS94a       0.4450      0.4577      1.0000
                    NS92a       SS92a       SS94a
```

The Akaike information criteria (AIC) and Schwarz information criteria (SIC) can be used to compare non-nested models (for details see Section 10.4). Smaller values of AIC or SIC indicate a better fit. The difference between the two criteria is that SIC takes into account the number of observations in the data set, and therefore will usually select a more parsimonious model than AIC. The Summary table displayed in the output is sorted in increasing order based on our selected criterion (SIC). SIC has identified FA (Factor analytic, order 1) as the "*best model*" for yld (SIC = 17524). FA2 (Factor analytic, order 2), a less parsimonious model, has the smallest AIC.

The residual variance model and variance-covariance matrix are outputted for the "*best model*", FA. The residual variance model displays the estimates and standard errors for each parameter. In this example, there are eight factor 1 gammas (g_11, g_21, …) and eight specific variances (psi_1, psi_2, …), one for each environment. The variance-covariance matrix is ordered alphabetically on env level names, indexed 1 to 8 (1=HN96b, 2=IS92a, …). Using this matrix we can identify environments with small and large genetic variances, and the explore relationships between environments. For example, the intermediate stress environments IS92a (2) and IS94a (3) have genetic variances of 21373 and 21860, respectively and a genetic covariance of 12558. This corresponds to a genetic correlation of 0.58. A high genetic correlation occurs when all genotypes are responding similarly to environmental differences. Conversely, a low genetic correlation arises when genotypes react very differently, indicating a strong G×E interaction.

4.3   Accounting for within-trial plot variation

In stage I of the two-stage analysis process for modelling MET data, the trait means for each genotype are generated separately for each trial (see Chapter 3). The stage II analysis should be "*weighted*" to accommodate both heterogeneity of error variance across trials and unequal replication within trials. This is achieved via the quantity e_{ij} in Equation 1; the error in the data arising from within-trial variation, which is referred to in Genstat as the "*unit error*". The unit errors are a measure of precision of the trait means and their inclusion in G×E analysis can improve genotype predictions when trials within a MET have different levels of precision (Welham *et al*., 2010; Möhring & Piepho, 2009).

Genstat follows the approach of Smith *et al.* (2001a) to generate unit errors from raw plot data using the Preliminary Single Environment Analysis menu (see Section 3.2 for

details). Other weighting schemes have been proposed (see Möhring & Piepho, 2009) but are not yet available under the QTL menu.

A good choice of weights will produce results from a two-stage analysis very similar to those from a one-stage analysis, in which plot data are analysed instead of means (see Smith *et al.,* (2005) for a description of the one-stage approach). There are theoretical advantages of the one-stage analysis over the two-stage analysis (Welham *et al.,* 2010), however the two-stage approach is logistically and computationally much easier to manage.

In this section we illustrate a weighted stage II analysis using trait means and unit errors from the 4 environment CIMMYT spring wheat trials (Section 1.3.2) held in Genstat spreadsheet. `SxBmeans.GSH`. Open the Select Best Variance-covariance Model menu (see Figure 4.3) and set Quantitative trait means: to `yield`, Genotype factor: to `genotype` and Environment factor: to `environment`. We will select all available variance-covariance models, and compare their goodness of fit using the Schwarz information criterion (SIC). As the MET comprises < 5 environments, the `FA2` (Factor analytic, order 2) model is unavailable.

To include unit errors, e_{ij}, in the analysis open the Select Variance-covariance Model Options window and check Include unit error: (Figure 4.5). If the unit error data structure has been stored in the QTL Data Space it will be automatically entered into the input field.
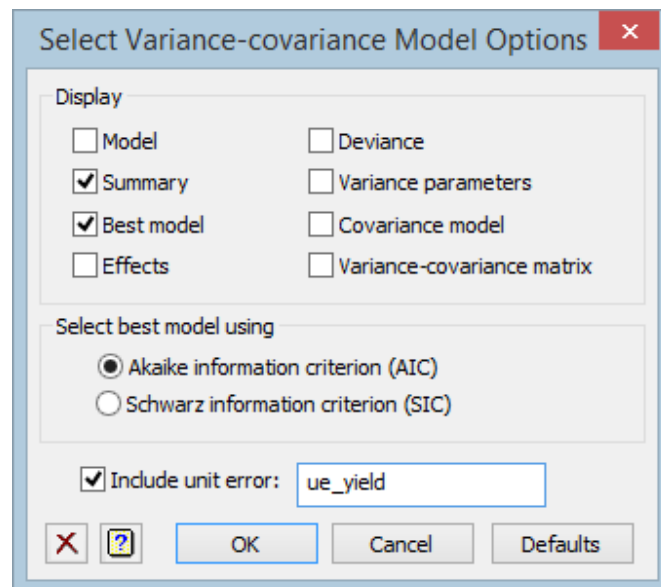
Figure 4.5: Using the Select Best Variance-covariance Model Options window to include unit errors (`ue_yield`), representing uncertainty in trait means (`yield`), in the G×E analysis of the CIMMYT spring wheat trials.

The output from the weighted analysis of `yield` is shown below:

```
Summary
=======


Trait: yield

          Model          AIC          SIC     Deviance NParameters
             CS         5231         5237         5227           2
        OUTSIDE         5224         5239         5214           5
            HCS         5226         5241         5216           5
             FA         5224         5249         5208           8
   UNSTRUCTURED         5225         5256         5205          10
       IDENTITY         5515         5518         5513           1
       DIAGONAL         5513         5525         5505           4


Best model: CS (on basis of criterion SIC)



Estimated variance components
-----------------------------

Random term                 component        s.e.
genotype                      991.341     127.263
genotype.environment          334.164      39.076



Residual variance model
-----------------------

Term                        Model(order)  Parameter      Estimate      s.e.
qtl_unitfactor                  Identity     Sigma2          1.000     fixed


Correlation matrix:
        DRIP05     1.0000
        HEAT06     0.5798       1.0000
        IRRI06     0.5488       0.5860       1.0000
        HEAT05     0.6154       0.6571       0.6220       1.0000
                   DRIP05       HEAT06       IRRI06       HEAT05
```

Schwarz information criteria (SIC) selects the `CS` (Compound Symmetry) variance-covariance model, which generates a uniform covariance structure, with equal variances and covariances across environments. That is, the variation between genotype effects is the same within each environment, and the covariance across environments is the same for every pair of environments.

We can compare this to an unweighted analysis, by unchecking Include unit error:. In this case `CS` is still selected as the best variance-covariance model. However, when the information on precision from the stage I analysis is not used, the residual variance term cannot reliably be separated from any uncorrelated component of the G×E effects, u_{ij}. This is equivalent to fitting a main effects model, with `environment` as fixed and `genotype` as random, and allocating their interaction to the residual term.

```
Summary
=======

Trait: yield

        Model        AIC        SIC   Deviance NParameters
           CS        5221       5227       5217           2
      OUTSIDE        5224       5240       5214           5
          HCS        5227       5242       5217           5
           FA        5225       5250       5209           8
 UNSTRUCTURED        5225       5256       5205          10
     IDENTITY        5510       5513       5508           1
     DIAGONAL        5513       5525       5505           4

Best model: CS (on basis of criterion SIC)


Estimated variance components
-----------------------------

Random term                  component        s.e.
genotype                         960.4       123.5


Residual variance model
-----------------------

Term                     Model(order)  Parameter        Estimate      s.e.
genotype.environment        Identity      Sigma2            618.4      39.4


Correlation matrix:
      DRIP05      1.0000
      HEAT06      0.6083      1.0000
      IRRI06      0.6083      0.6083      1.0000
      HEAT05      0.6083      0.6083      0.6083      1.0000
                  DRIP05      HEAT06      IRRI06      HEAT05
```

4.4 Exploratory methods for G×E interaction

Genstat provides two descriptive tools for exploring the G×E interaction; AMMI models and GGE biplots. They are accessed by selecting the Phenotypic Data option under the Data Exploration menu. We'll look at both of these in turn.

4.4.1 AMMI

MET data can be classified by two factors (genotype and environment). An Additive Main Effects and Multiplicative Interaction (AMMI) model is a hybrid analysis that incorporates both the additive and multiplicative components of the two-way MET data structure. A detailed overview of the AMMI methodology can be found in Crossa and Cornelius (2002). In brief, the AMMI analysis extracts genotype and environment main effects (i.e. the additive component) using analysis of variance (ANOVA). Next, principal components analysis (PCA) is applied to the ANOVA residuals, which include the G×E interaction (i.e. the multiplicative component). The PCA partitions the G×E interaction into IPCA (I for interaction) components. The first component explains the most variation in the G×E interaction, followed by the second, and so on. Each IPCA component is the product of a genotypic and an environmental score. These scores can be used to construct biplots (Gower and Hand, 1996), a powerful graphical representation that help us explore the G×E interaction.

Genstat's AMMI procedure accommodates both raw plot (unit) data and trait means. However, the data must be balanced. That is, the same genotypes at each environment and, in the case of the plot data, the same number of replicates for each genotype by environment combination. In addition, there must be no missing values. When analysing plot data with an unbalanced or complicated randomization structure, it is recommended that you form the genotype by environment trait means first (see Section 3.2) and supply these instead.

We shall demonstrate AMMI modelling using `yld` trait means from the CIMMYT maize trials (Section 1.3.2) held in file `F2_maize_pheno.csv`. A detailed discussion of this example can be found in Malosetti *et al.* (2013).

The AMMI menu (Figure 4.6) can be accessed via Stats | QTLs (Linkage/Association) | Data Exploration | Phenotypic Data | AMMI; or, by using the Explore button in the QTL Data View. The Data:, Genotypes:, and Environments: fields are automatically filled using data in the QTL Data Space. Set the Data: field to `yld`.
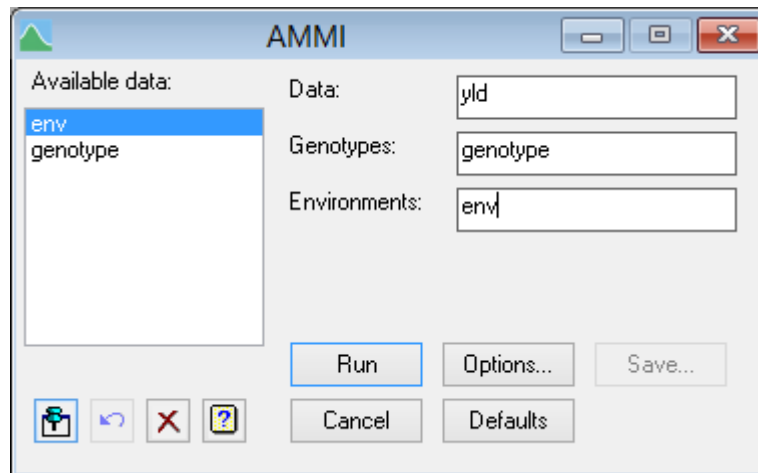
Figure 4.6: AMMI window for `yld` trait means from the `F2maize_pheno.csv` data set.

The Options button opens a window that allows you to select the output and graphs to be generated by an AMMI analysis (Figure 4.7).
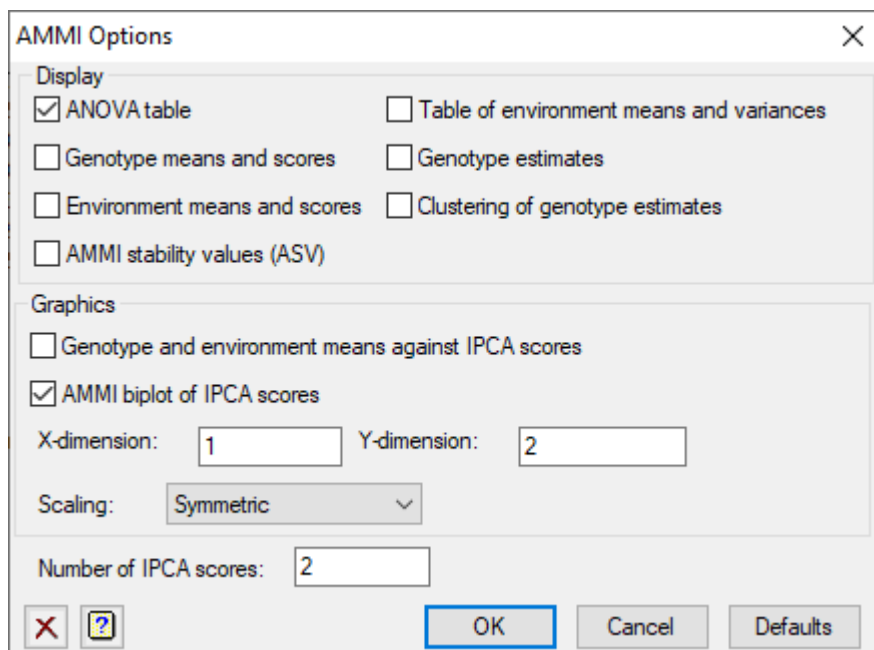
Figure 4.7: AMMI Options window.

We select to display the ANOVA table, and to graph Genotype and environment means against IPCA scores and the AMMI biplot of IPCA scores. The set of IPCA scores to plot are specified using X-dimension: and Y-dimension: fields. Here, we request the first set is

plotted on the x-axis, and the second set on the y-axis. The Number of IPCA scores:, which defines the number of IPCA components (or sets of genotypic and environmental scores) used to partition the G×E interaction, defaults to 2. The Scaling: option controls the scaling used for the biplot. Genotype-focused scaling (Scaling: = Genotype) is best suited to displaying the inter-relationships among genotypes, whereas the environment-focused scaling (Scaling: = Environment) is better suited to displaying the inter-relationships among environments. The Symmetric scaling method is an intermediate between environment-focused and genotype-focused scaling.

Once the analysis is run the Save button (Figure 4.6) is activated allowing the genotype and environment IPCA scores, and the fitted values and residuals from the AMMI model to be saved.

The ANOVA table, which summaries the contribution of each IPCA component to the interaction term, is given below. We have partitioned the interaction term into two components (IPCA 1 and IPCA 2). Both are highly significant (p-value<0.001), explaining 29.8% (5451796/18296997×100%) and 21.3% (3888148/18296997×100%) of the G×E interaction sum of squares, respectively.

```
AMMI Analysis
=============


ANOVA table for AMMI model
--------------------------


Source          d.f.        s.s.        m.s.        v.r.        F pr
Genotypes        210    13821018       65814        5.29      <0.001
Environments       7   127771687    18253098     1466.47      <0.001
Interactions    1470    18296997       12447
  IPCA 1         216     5451796       25240        2.93      <0.001
  IPCA 2         214     3888148       18169        2.11      <0.001
  Residuals     1040     8957053        8613
```

A biplot helps us to visualize the G×E interaction, and to identify genotypes with broad (or specific) adaptability and environments which elicit strong (or weak) interactive forces. The biplot for the yld data, using symmetric scaling and the first two IPCA components, is given in Figure 4.8. Genotypes are represented by green crosses (×), and environments by blue pluses (+), with vectors connecting the environment scores with the origin. Genotypes that cluster together behave similarly across the environments (e.g.

`G061` and `G047`), whilst environments that cluster together influence the genotypes in a similar way (e.g. `LN96a` and `LN96b`). The angle between the environment vectors provides further information on the correlation between environments, where;

- an acute angle indicates positive correlation (e.g. `LN96a` and `LN96b`)
- an obtuse angle indicates negative correlation (e.g. `HN96b` and `IS92a`)
- a right angle indicates no correlation (e.g. `HN96b` and `NS92a`)
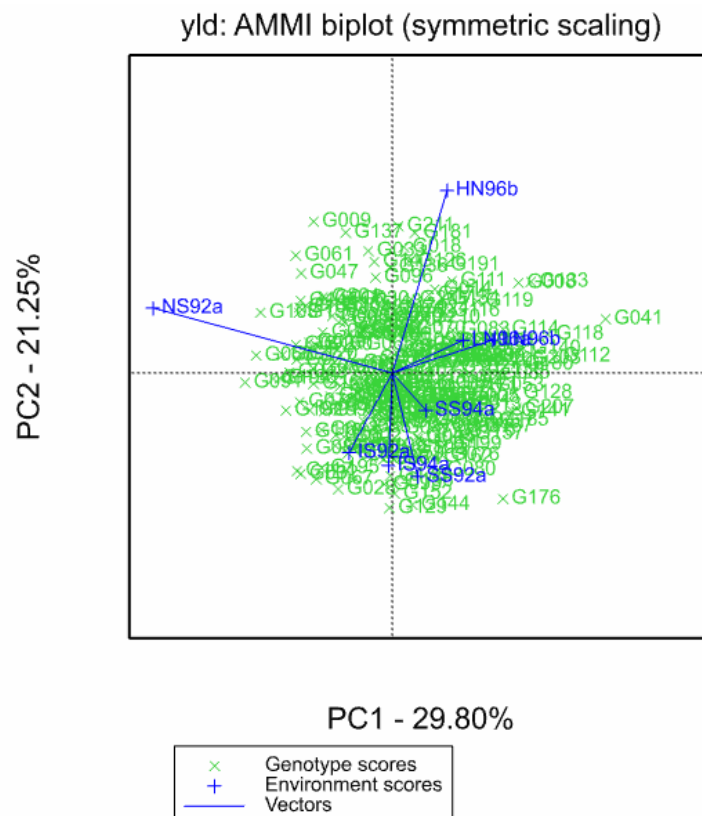
Figure 4.8: Biplot from the AMMI model for `yld`.

The biplot's origin represents the overall mean phenotypic response, in our case of `yld`. The position of a genotype, or environment, relative to the origin provides insights into the nature of the G×E interaction. In general, genotypes which are near the origin are insensitive to environmental interactions, i.e. are broadly adapted (whether good or poor performance depends on the genotype main effect). Conversely, genotypes far from the origin are sensitive to environmental interactions, i.e. are specifically adapted (e.g. `G176`). Likewise, environments near the origin (as indicated by a short vector) elicit only weak

interactive forces (e.g. SS94a) whereas those far from the origin elicit strong interactive forces (e.g. NS92a).

As the IPCA axes explain a portion of the G×E interaction, some caution must be taken when interpreting the biplot. For example, a genotype showing a particularly unique interaction with environment will plot at the origin if it is explained by a higher order IPCA axis.

Plots of the genotype and environment means against the first (Figure 4.9a) and second (Figure 4.9b) IPCA scores are useful for identifying genotypes that have broad and/or specific adaptability across the environments studied. To interpret these plots we examine differences along the x-axis, which indicate differences in the main (additive effects), and distances from zero on the y-axis, which indicate the strength of the G×E interaction. Genotype G041 stands out in Figure 4.9a. This genotype has a low mean yld, but interacts strongly with the environments. A positive interaction (i.e. improved relative performance) is indicated by a genotype and an environment having the same sign on the IPCA axis (y-axis). Conversely, different signs imply a negative interaction (i.e. reduced relative performance). Genotype G041 has a different sign to environment NS92a and the same sign as environment LN96b (Figure 4.9a). This indicates that it performs relatively poorly in environment NS92a but relatively well in environment LN96b.
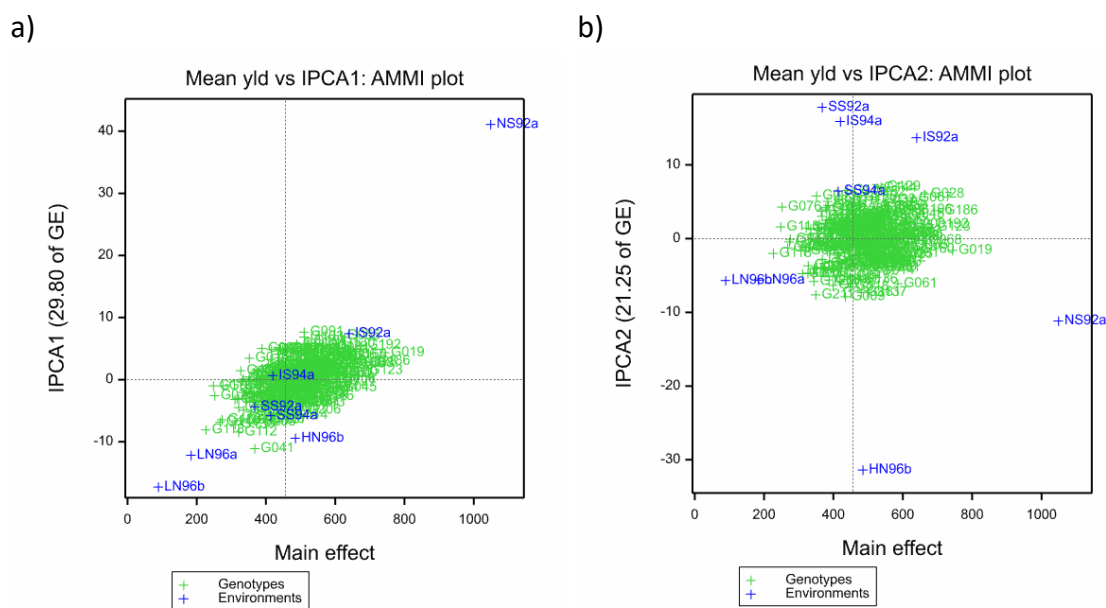
Figure 4.9: Plot of the genotype and environment means against the a) first and b) second IPCA scores.

4.4.2 GGE biplot

Genotypic main effect plus Genotype-by-Environment interaction (GGE) biplots are a useful tool for assessing the phenotypic performance of genotypes in different environments. The GGE biplot, a modification of the AMMI model, provides information on total genetic variation by approximating the joint effect of the genotype and G×E interaction. In contrast, AMMI biplots, discussed in Section 4.4.1, approximate only the G×E interaction component of genetic variation. In brief, the GGE biplot is based on a standard PCA of the environment-centred trait means. The GGE biplot summarizes the genotype plus G×E variation using scores from the PCA. A comprehensive description of the methodology can be found in Yan and Kang (2003). The interpretation of the GGE biplot is very similar to the AMMI biplot. However, now the genotypes are distributed according to their overall phenotypic performance in each environment, rather than by the size of the G×E interaction.

The GGE biplot menu can be accessed via Stats | QTLs (Linkage/Association) | Data Exploration | Phenotypic Data | GGE Biplot; or, by using the Explore button in the QTL Data View. The analysis uses G×E trait means (refer to Section 3.2 for information on how to generate these). We demonstrate the use of GGE biplots using trait means from CIMMYT maize trials (Section 1.3.2) held in file `F2_maize_pheno.csv`.

The GGE Biplot menu is automatically populated using available data in the QTL Data Space (Figure 4.10). We choose to analyse `yld`. The Type of plot: field defaults to Scatter plot, which is the standard PCA biplot.
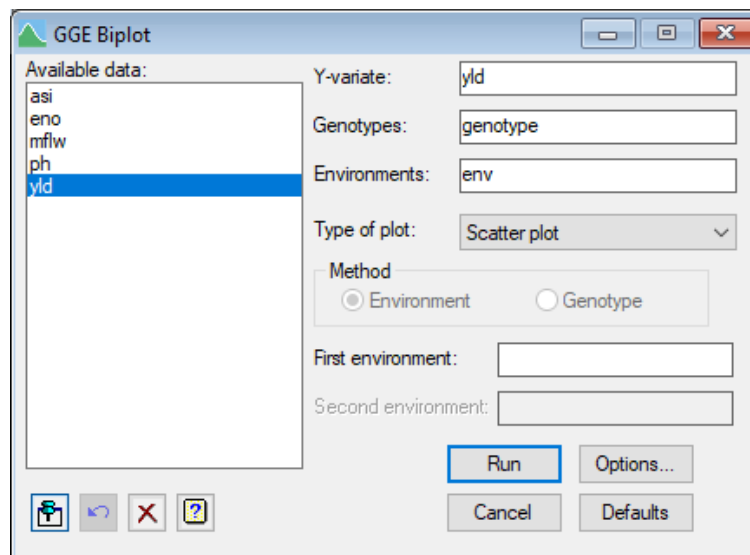
Figure 4.10: GGE Biplot window for `yld` trait means from `F2maize_pheno.csv`.

The Options button opens a menu window were additional options and settings can be specified (Figure 4.11). The Scaling option defaults to Environment. This is recommended, together with Connect environment scores with origin, for displaying the inter-relationship among environments. Complete information on all options is available by clicking on the help icon (⯑).
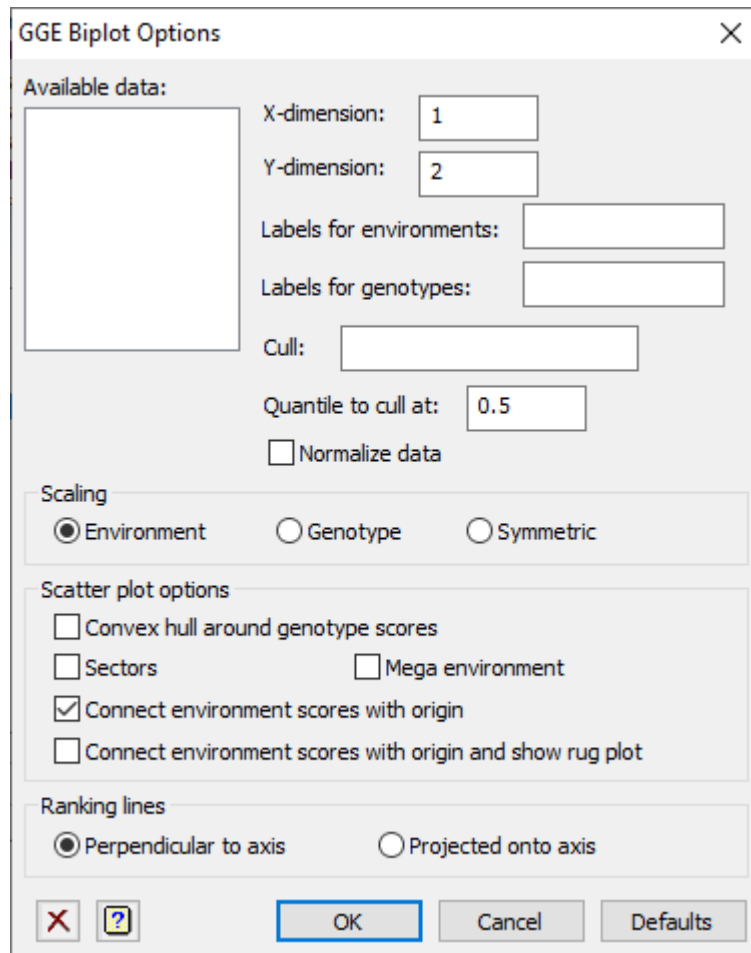
Figure 4.11: The GGE Biplots Options window.

The features of a GGE biplot (Figure 4.12) have a similar interpretation to that of an AMMI biplot (Section 4.4.1). For example, genotypes (or environments) that are alike cluster together. The key difference being, the genotypic scores now jointly describe both the genotypic main effect and the G×E interaction effect. Therefore, the highest yielding

genotypes, such as `G019`, are found on the right-hand side of the biplot, and the lowest yielding, such as `G118`, on the left.
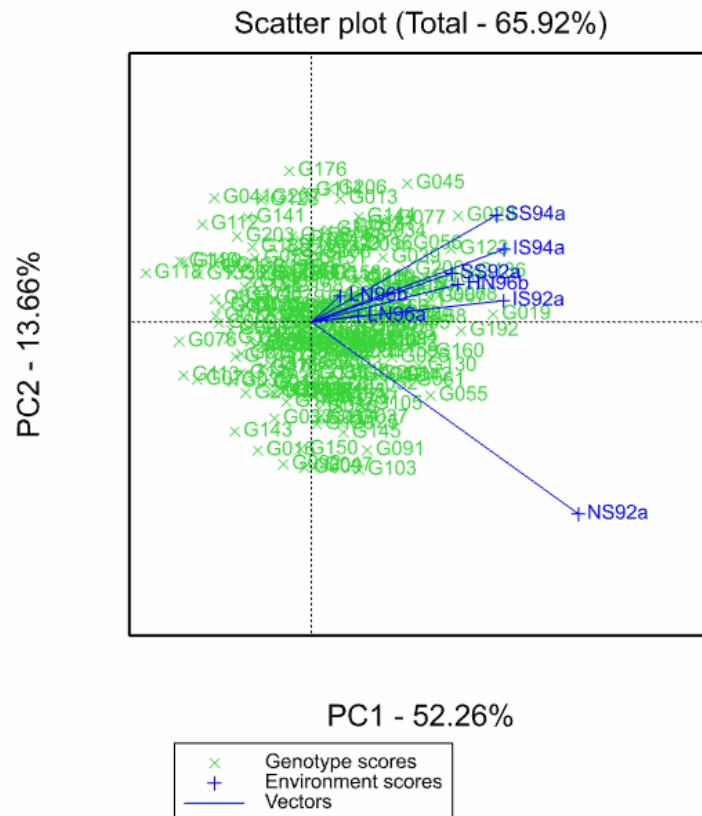
Figure 4.12: GGE biplot for `yld`.

Additional information, that may help elucidate the genotype and environment relationships, can be added to Figure 4.12 by changing the Type of plot:, Method, First environment: and Second environment: fields in the GGE Biplot menu (Figure 4.10). Detailed information on these options can be found in the GGE biplot help (accessed via the help icon ▣) and in the documentation for the GGEBIPLOT procedure. The different types of GGE biplot are:

- Scatter plot - plots the genotype and environment scores.
- Ranking plot - shows best performing genotypes (or environments) in a specific environment (or genotype).
- Comparison - compares the performance of the environments (or genotypes) with that of an "*ideal*" environment (or genotype).
- Joint - compares two environments (or genotypes) simultaneously.

- Centred scatter plot - compares the performance of the genotypes (or environments) in two environments (or genotypes).

For example, the GGE biplot used to illustrate which genotypes perform the best (i.e. highest yielding) in environment `NS92a` is generated by setting Type of plot: to Ranking plot and specifying the environment of interest (either its level, or label within single quotation marks) in the First environment: field (Figure 4.13). This draws a biplot axis through our specified environment (`NS92a`) together with ranking lines to show the best performing genotypes in this environment (Figure 4.14).
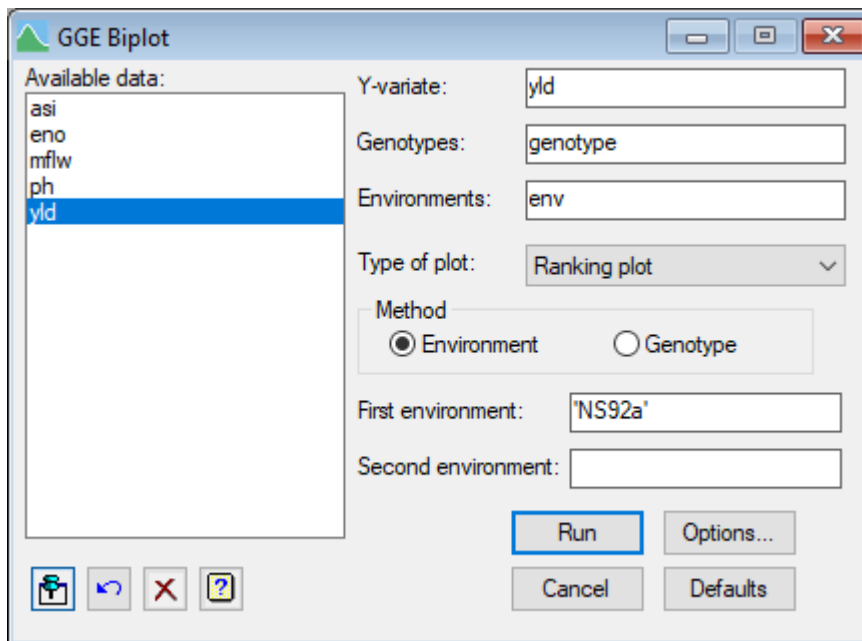
Figure 4.13: Menu options to produce a ranking biplot for environment `NS92a`.

Using Figure 4.14 we can readily compare the performances of all genotypes within environment `NS92a`. The genotypes are "*projected*" onto the biplot axis (which runs along the vector for `NS92a`) using perpendicular lines. The best performing genotypes in environment `NS92a` are those whose projections onto the biplot axis are closest to the score for `NS92a`. By default the ranking lines are drawn to be perpendicular to the biplot axis (as in Figure 4.14), but you can project lines from the genotypes to the biplot axis by setting Ranking lines to Projected onto axis in the GGE Biplot Options window (see Figure 4.11). The best performing genotypes in environment `NS92a` are `G019`, `G055`, `G103`, `G192` and `G091`. The poorest performing genotypes are `G118`, `G041` and `G112`.
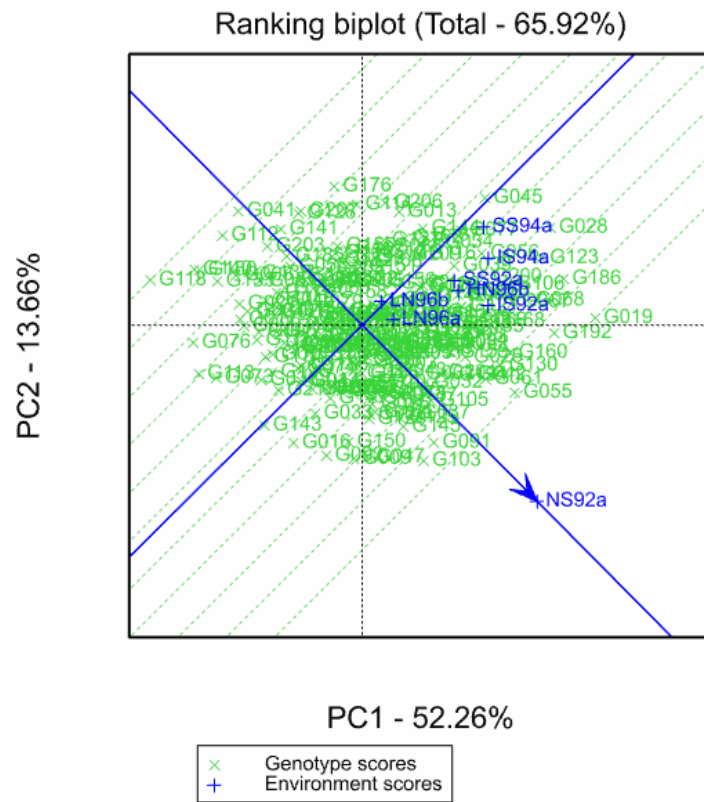
Figure 4.14: Ranking biplot of genotypes in environment `NS92a`.

4.5   References

Crossa, J., & Cornelius, P.L. (2002). Linear-bilinear models for the analysis of genotype-environment interaction. In *Quantitative genetics, genomics and plant breeding*. CAB International, New York. 305-322.

Falconer, D.S., & Mackay, T.F.C. (1996). Introduction to quantitative genetics. 4th edition. London: Longman Scientific and Technical.

Gower, J.C., & Hand, D.J. (1996). Biplots. Chapman & Hall, London.

Malosetti, M., Ribaut, J.-M., & van Eeuwijk, F.A. (2013). The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Frontiers in Physiology*, **4**, 44.

Möhring, J., & Piepho, H.-P. (2009) Comparison of weighting in two-stage analysis of plant breeding trials. *Crop Science*, **49**, 1977-1988.

Patterson, H.D., & Silvey, V. (1980). Statutory and recommended list trials of crop varieties in the United Kingdom. *Journal of Royal Statistical Society*, *A,* **143**, 219-252.

Smith, A., Cullis, B., & Gilmour, A. (2001a). The analysis of crop variety evaluation data in Australia. *Australian and New Zealand Journal of Statistics*, **43**, 129-145.

Smith, A., Cullis, B., & Thompson, R. (2001b). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics*, **57**, 1138-1147.

Smith, A.B., Cullis, B.R., & Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *Journal of Agricultural Science*, **143**, 449–462.

Welham, S.J., Gogel, B.J., Smith, A.B., Thompson, R., & Cullis, B.R. (2010). A comparison of analysis methods for late-stage variety evaluation trials. *Australian & New Zealand Journal of Statistics*, **52**, 125-149.

Yan, W., & Kang, M.S. (2003). GGE biplot analysis: a graphical tool for breeders, geneticists and agronomists. CRC Press, Boca Raton.

5 Construction of genetic linkage maps

QTL analysis requires that a genetic linkage map is available for the population under study. If that is not the case, Genstat provides facilities for constructing a map from a set of marker scores. These facilities are accessed via Stats | QTLs (Linkage/Association) | Map Construction (Figure 5.1); or, by using the Map construction button in the QTL Data View.
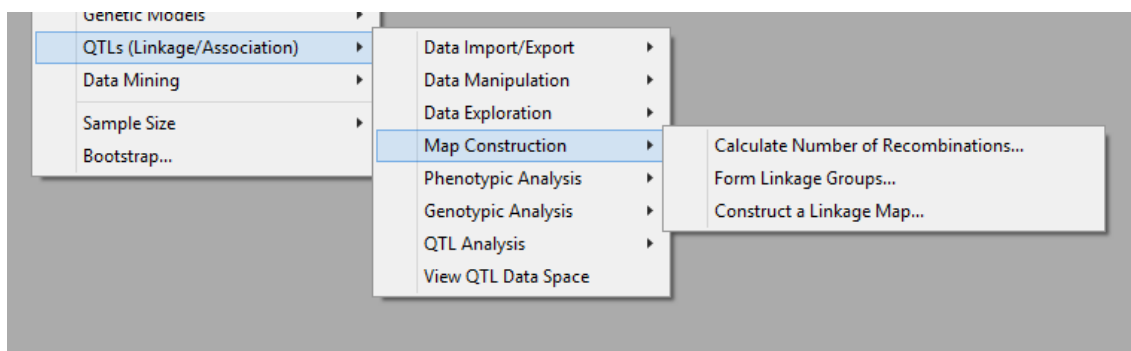
Figure 5.1: Map construction facilities.

5.1 Number of recombinations

The Calculate Number of Recombinations menu (Figure 5.2) is used to estimate the number of recombinations and the recombination frequencies between markers. Input fields Marker genotype scores:, Marker names: and Parental information: are automatically populated using available structures in the QTL Data Space. The appropriate population type (F2, DH1, BC1, RILn or CP) needs to be specified in the Type of population: field.

The number of recombinations can be estimated using the Two point method (see Maliepaard *et al*., 1997), or, when marker order is available, the Multi-point method (see Lander & Green, 1987). If Multi-point is checked, the field Initial order of markers: activates where a variate containing the marker order is supplied.

Clicking on the Store button opens a window allowing you to save results from analysis. (For the two point method: expected number of recombinations between markers and estimated recombination frequencies. For the multi-point method: position of markers, inheritance vectors and expected number of recombinations of the

genotypes.) After checking the appropriate boxes, names for the saved data structures need to be specified in the corresponding In: fields.

The Options button opens a window where you can specify the output to produce. For the two point method, you can check Increase number of recombinations by 0.5 for each missing value to impose a penalty for missing data. The number of recombinations is then increased by 0.5 recombination per informative meiosis for each missing marker score.
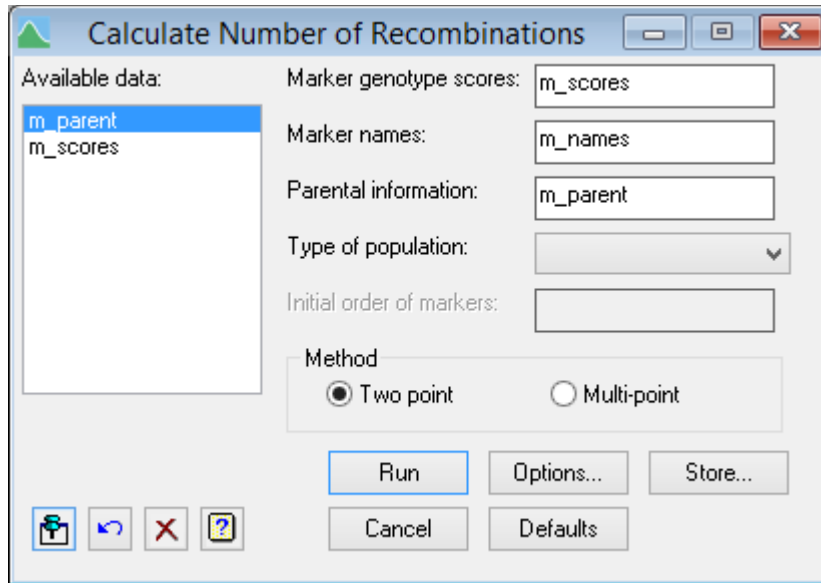
Figure 5.2: Calculate Number of Recombinations menu.

5.2 Formation of linkage groups

The Form Linkage Groups menu (Figure 5.3) is used to form linkage groups from marker data. The population (F2, DH1, BC1, RILn or CP) is specified in the field Type of population:. The other input fields are automatically populated using available data in the QTL Data Space.

The analysis first calculates recombination frequencies, using the two point method, then linkage groups are formed using depth-first search from a symmetric matrix of links. The threshold for the recombination frequency at which markers are said to be linked (default 0.2) can be specified in the Form Linkage Groups Options menu (click on the Options button to open the options menu).

In the Form Linkage Groups Options menu you can also request a Summary of the number of markers in each linkage group, and specify whether to Increase number of recombinations by 0.5 for each missing value.

Figure 5.3: Form Linkage Groups menu.

5.3   Construct genetic linkage maps

The Construct a Linkage Map menu (Figure 5.4) is used to calculate the order and position of the markers on a chromosome (or linkage group). Using data available from the QTL Data Space, the Marker genotype scores:, Marker names: Genotype labels: and Parental information: fields are automatically populated. The population type (F2, DH1, BC1, RILn or CP) must be specified in the Type of population: field.
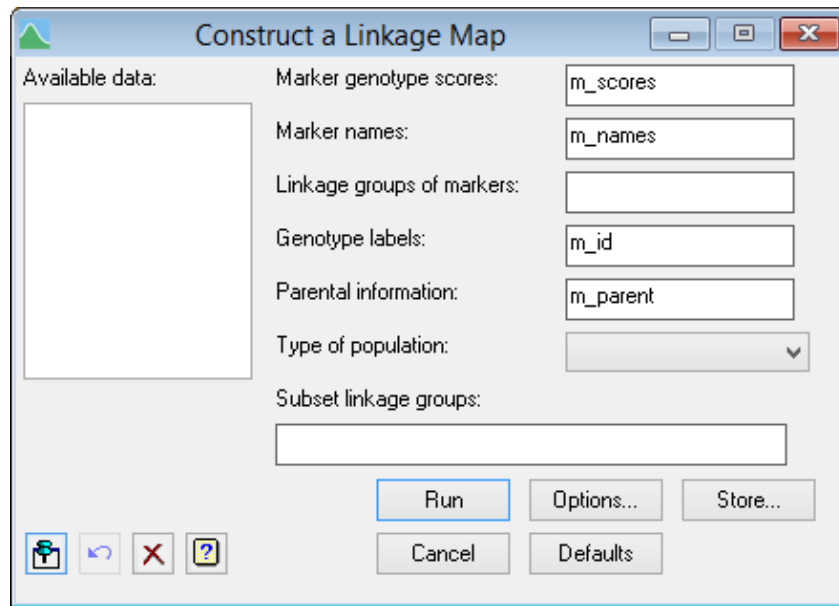
Figure 5.4: Construct a Linkage Map menu.

All markers are assumed to belong to the same linkage group unless a factor defining the linkage groups for each marker is specified in Linkage groups for markers:. If this is set, the marker positions are calculated within each level of the linkage group factor. Further, the Subset linkage groups: field can be used define a subset of linkage groups to map.

The analysis calculates the order of markers using simulated annealing in conjunction with spatial clustering; either spatial sampling or optimization (for more information see Lander & Green, 1987; Jensen *et al*., 2001; Jensen, 2005). Spatial clustering is used to obtain a framework map: this reduces the size of the optimization problem and leads to a reduction of the effects of errors on the marker ordering. You can select the type of spatial clustering used to obtain a framework map in the Construct a Linkage Map Options window (Figure 5.5), launched by clicking the Options button.
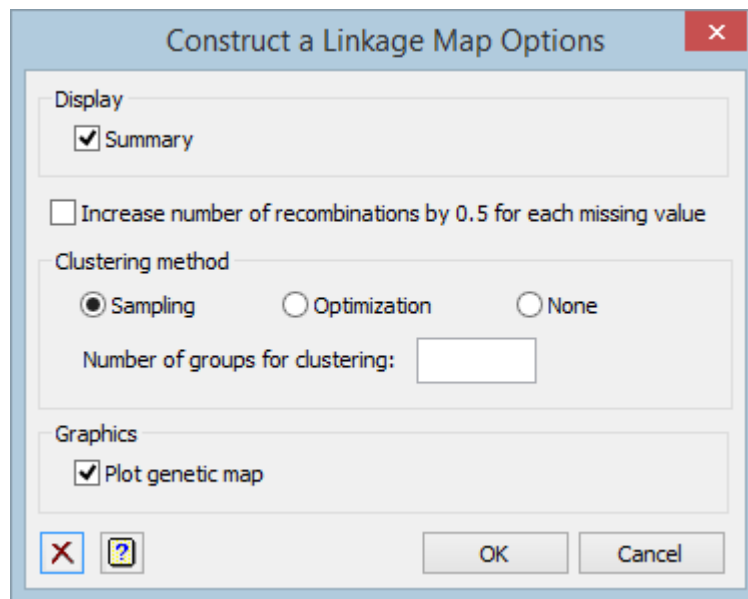
Figure 5.5: Options for constructing a linkage map.

The Construct a Linkage Map Options menu also allows you to request summary information (i.e. number of linkage groups and the minimum, mean and maximum of the positions per linkage group) and a plot of the genetic map.

Clicking on the Store button opens a window allowing you to save results from analysis.

5.4   References

Jansen, J., de Jong, A.G., & van Ooijen, J.W. (2001). Constructing dense genetic linkage maps. *Theoretical and Applied Genetics*, **102**, 1113-1122.

Jansen, J. (2005). Construction of linkage maps in full-sib families of diploid outbreeding species by minimizing the number of recombinations in hidden inheritance vectors. *Genetics*, **170**, 2013-2025.

Lander, E.S., & Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 2363-2367.

Maliepaard, C., Jansen, J., & van Ooijen, J.W. (1997). Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. *Genetics Research*, **70**, 237-250.

6   Linkage analysis: inbred population with a single trait evaluated at a single site

Linkage analysis is a method for detecting quantitative trait loci (QTLs). It tests whether variation at the molecular level is related to phenotypic variation in a specifically designed segregating population.

Most traits plant breeders work with are quantitative. What complicates the breeding task is that trait variation is the result of a large number of QTLs, each one with a minor effect. The objective of linkage analysis is to dissect the complexity of quantitative traits by identifying the individual QTLs. A typical QTL experiment consists of developing a segregating population (F2, double haploid, back-cross, recombinant inbred lines, back-cross inbred lines or F1 cross-pollinating) and finding statistical associations between the observed phenotypic variation and variation at the DNA level (measured using molecular markers).

In this chapter, linkage analysis is illustrated for an inbred population with a single trait evaluated at a single site (i.e. in a single "*environment*"). Linkage analysis for more complex data sets will be described in Chapters 7 and 8.

In this chapter you will learn how to detect QTLs in a single trait - single environment data set, including how to:

- calculate genetic predictors for use in linkage analysis (Section 6.1.1)
- perform marker regression (Section 6.1.2.1)
- perform a simple interval mapping (SIM) scan (Section 6.1.2.2)
- select co-factors and perform a composite interval mapping (CIM) scan (Section 6.1.2.3)
- determine a final multi-QTL model (Section 6.1.2.4)
- correct for multiple comparisons (Section 6.1.4)

6.1 QTL linkage analysis

The aim of linkage analysis is to identify QTLs by linking phenotypic variation with molecular variation. The term "*linkage*" arises from the proximity, or degree of linkage, between the marker and the detected QTL.

Conceptually we can think of QTL detection as a model selection problem, with the purpose of selecting a model that describes the phenotypic response in terms of QTL effects. The process of detecting QTLs can be divided into two major parts: a) a putative QTL detection step, where the genome is searched for candidate QTLs, and b) selection of the final multi-QTL model from the set of candidate QTLs.

The Genstat QTL menu is tailored for detecting QTLs in bi-parental plant breeding populations. It is designed to guide the user through a step-by-step strategy for QTL detection. These steps comprise of:

1) calculating genetic predictors, which are used as explanatory variables in the QTL models (Section 6.1.1)
2) an initial genome-wide scan by marker regression or simple interval mapping (SIM) to obtain candidate QTL positions (Sections 6.1.2.1 and 6.1.2.2, respectively)
3) one or more rounds of composite interval mapping (CIM), in which co-factors correct for QTLs that segregate elsewhere in the genome (Section 6.1.2.3)
4) fitting a final multi-QTL model, usually following back-selection from a set of candidate QTLs, to get a final set of estimated QTL effects (Section 6.1.2.4).

This chapter illustrates the QTL detection process for the simplest type of data: a single trait, single environment data set. Yield data from the Steptoe-Morex barley trial (introduced in Section 1.3.1) is used as an example throughout.

6.1.1 Calculation of genetic predictors

Prior to QTL analysis, genetic predictors are calculated to be used as explanatory variables in the QTL models. Genetic predictors are constructed from molecular marker information, and QTLs are identified by genetic predictors that have some explanatory power for the phenotypic trait of interest.

In Genstat, genetic predictors can be formed from allele frequencies. For example, the "*additive*" genetic predictor takes the value -1 when an individual is homozygous like parent 1 (P_1P_1), 0 when heterozygous (P_1P_2), and 1 when homozygous like parent 2 (P_2P_2). The slope from a simple linear regression of the trait means on the additive genetic

predictor represents the effect of replacing an allele from parent 1 (P_1) with an allele from parent 2 (P_2), known as an "*additive genetic effect*". Deviations from additivity can be explored by forming a "*dominance*" genetic predictor. Here, the predictor takes the value 0 when the individual is homozygous (i.e. P_1P_1 or P_2P_2), and value 1 when heterozygous (P_1P_2). As by definition dominance effects are deviations from additivity, testing for dominance is always done conditional on the additive effect being present in the model. For cross-pollinated populations, an additive genetic predictor for the 2nd parent can also be formed.

In practice, marker data frequently contains missing values, especially when the markers are dominant. Constructing genetic predictors for incomplete marker data, or at genomic positions in between marker loci, is more complex (see Jiang and Zeng, 1997). The methodology uses information from nearby markers to estimate the probability an individual is of a particular genotype. This probability depends on the observed genotypes at neighbouring markers and the distance from those markers. Genstat uses a Hidden Markov model to estimate these probabilities (which are known as "*conditional genotypic probabilities*"). The estimated probabilities are then used to obtain genetic predictors at genomic positions where no (or partial) marker information is available.

We will calculate genetic predictors for the Steptoe-Morex barley trial (Section 1.3.1) using Genstat. The marker and map information for this double haploid population are held in Flapjack files `SxM_geno.txt` and `SxM_map.txt`, respectively. Import and inspect the marker and map data (referring to Sections 2.1.2 and 2.4.2). Remember to specify the population as double-haploid (`DH1`).

On importing the genotypic data, Genstat automatically provides some summary information. This informs us that the Steptoe-Morex genotypic data set is from a double haploid population (`DH1`) containing 150 genotypes and 116 markers spread over 7 linkage groups. The number of markers per linkage group and the length of each linkage group is also provided.

```
Loading QTL data
----------------

Catalogue of files:
 C:/Program Files/Gen21Ed/Data/SxM_geno.txt
 C:/Program Files/Gen21Ed/Data/SxM_map.txt

Population:             DH1
Number of genotypes:    150
Number of markers:      116
Number of linkage groups: 7
```

```
Linkage group Number of markers        Length
            1                16         170.1
            2                17         183.1
            3                14         188.7
            4                12         167.0
            5                29         150.8
            6                13          95.5
            7                15         186.5


Parent IDs: Morex, Steptoe
```

The quality of the genotypic data, including the proportion and pattern of missing marker data, can be explored using the Data Exploration menu, as described in Section 2.4.2. We'll use the Stats | QTLs (Linkage/Association) | Data Exploration | Genotypic Data | Summary Statistics for Markers menu to output genotypes and markers with >10% missing values.

```
Missing values
--------------


There are 364 scores missing. This is 2.092% of the 17400 scores.


There are 60 markers with missing values. This is 51.72% of the 116 markers.

The 4 markers with more than 10% missing values over the 150 genotypes are:


                                  Number of       Percentage
  Marker Chromosome    Position missing values missing values
 abc310b           1       120.8            16           10.7
 bcd402b           4        33.7            17           11.3
  ksuh11           4       168.4            26           17.3
    aga6           5         0.0            67           44.7


There are 120 genotypes with missing values. This is 80% of the 150 genotypes.

The 2 genotypes with more than 10% missing values over the 116 markers are:

              Number of       Percentage
Genotype  missing values  missing values
   dh011              13            11.2
   dh012              12            10.3
```

Overall, 2% of the marker scores are missing. Four markers (`abc310b`, `bcd402b`, `ksuh11`, `aga6`) and two genotypes (`dh011`, `dh012`) have more than 10% missing. Marker `aga6`, in linkage group (chromosome) 5, has a very large number of missing values (45%) and perhaps should be omitted from the analysis.

The Genstat menu to calculate genetic predictors (Figure 6.1) can be accessed by Stats | QTLs (Linkage/Association) | Genotypic Analysis | Calculate Genetic Predictors; or, via the QTL Data View from Genotypic analysis | Calculate Genetic Predictors.
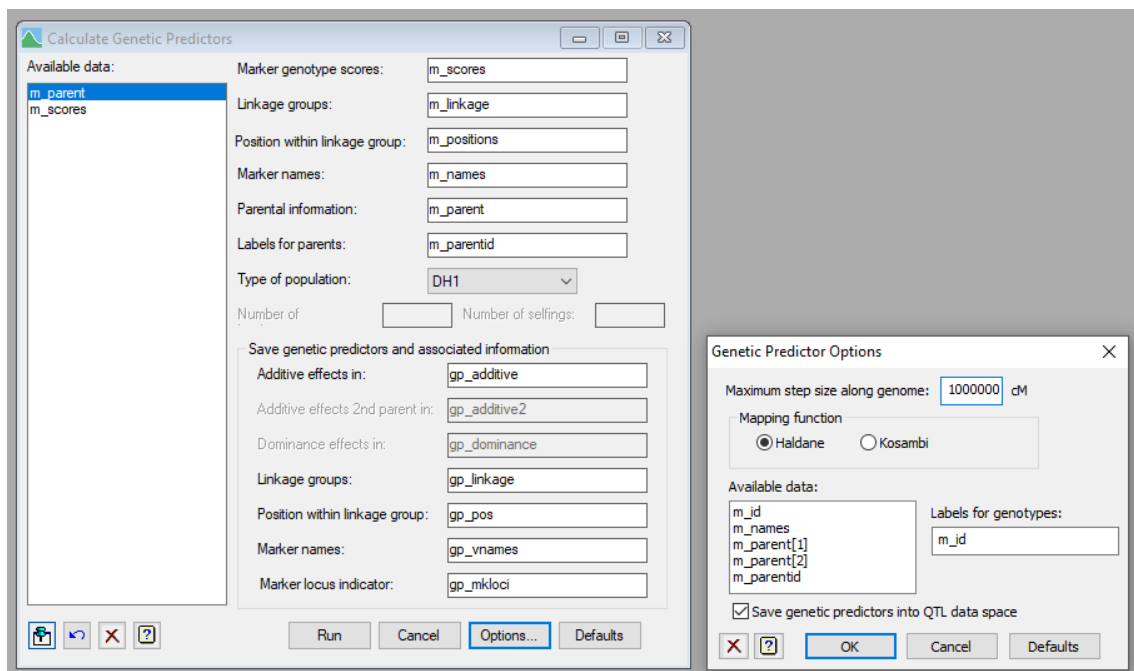
Figure 6.1: Menu and options for calculating genetic predictors.

Genstat will automatically populate the input fields from information in the QTL Data Space, and provide default names for saving the genetic predictors and their associated information. The fields greyed out in the Save genetic predictors and associated information box are unavailable for the Type of population: selected. For F2 populations, dominance effects can be obtained. For cross-pollinating (CP) populations, dominance effects and additive effects for the 2nd parent can be calculated.

Clicking on Options opens a window where you can specify the maximum step size between consecutive genetic predictors along the genome (Figure 6.1). The default of 10^6cM calculates genetic predictors at the marker positions only. Setting the maximum step size to some lesser value will calculate genetic predictors along the genome such that

the gap between any two consecutive predictors is less than this value. Using a maximum step size of 10cM as an example, if the distance between two consecutive genetic predictors is larger than 10cM, a new evaluation position will be created between them. The process is repeated until no gap of 10cM or more remains. For now, we will retain the default step size of 10^6cM.

Checking Save genetic predictors into QTL data space makes the names of the genetic predictor data structures available in subsequent QTL menus. Click OK and then Run to calculate the genetic predictors. Several new data structures will now appear in the *Genetic predictors* folder on the QTL Data View. To view the additive genetic predictors right click on `gp_additive` and select Create Spreadsheet (Figure 6.2).
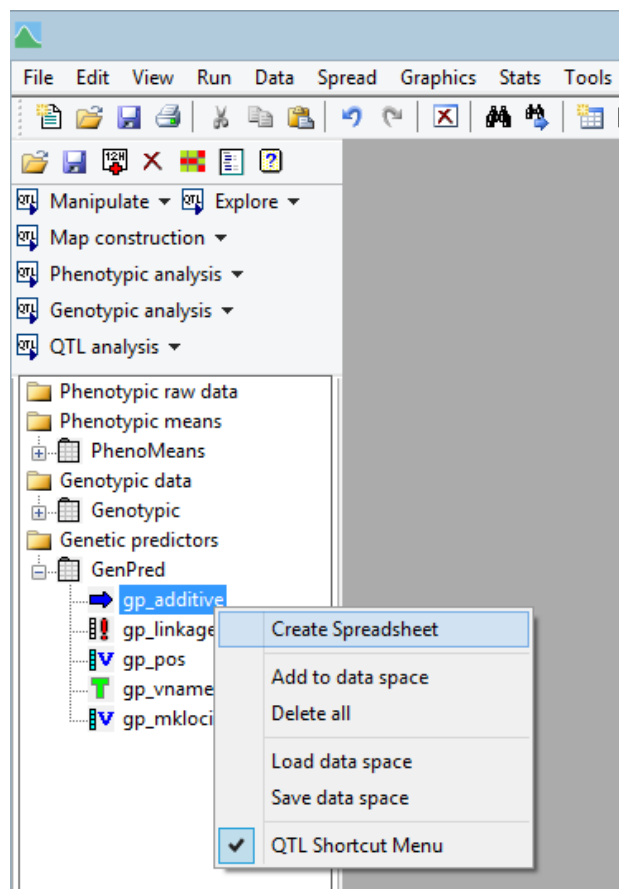
Figure 6.2: Creating a spreadsheet to display the additive genetic predictors.
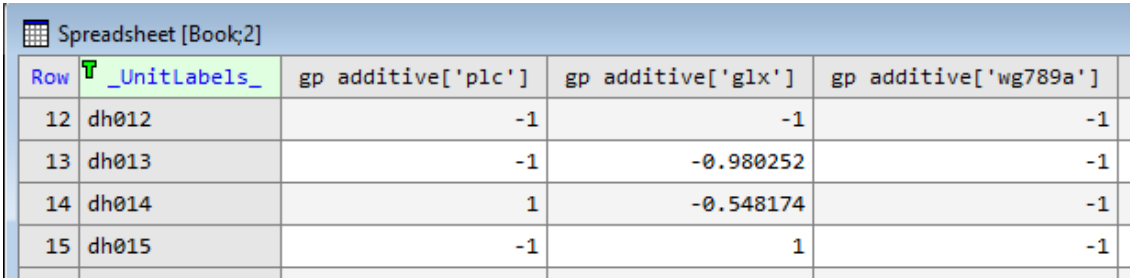
In Figure 6.3 is an excerpt of the Steptoe-Morex marker scores. Two scores are missing for the second marker (`glx`); for genotype `dh013` both neighbouring markers (`plc` and

`wg789a`) have a score of 1 (homozygous like parent 1), whilst for genotype `dh014` the neighbouring markers, `plc` and `wg789a`, have scores of 2 and 1, respectively.

```
LinkageGroup     1        1        1
Distance (cM)    0        18.7     24.1

Marker           plc      glx      wg789a
dh012            1        1        1
dh013            1        -        1
dh014            2        -        1
dh015            1        2        1
```

Figure 6.3: Steptoe-Morex marker scores for 4 genotypes from the first three markers on linkage group 1. A single code is used for homozygotes, where 1 = homozygous like parent 1, 2 = homozygous like parent 2. Missing values are indicated by -.

The values of the additive genetic predictors (`gp_additive`) corresponding to the marker data in Figure 6.3 are given in Figure 6.4. Homozygotes like parent 1 have been assigned a value of -1, homozygotes like parent 2 a value of 1, and missing values have been estimated. These estimates are produced using complex methodology based on estimating recombination frequencies with neighbouring markers (see Jiang and Zeng, 1997). However, intuitively, for genotype `dh013`, the estimate of `gp_additive['glx']`, -0.980, is close to -1, the value of both neighbouring genetic predictors, `gp_additive['plc']` and `gp_additive['wg789a']`. In contrast, `dh014` is estimated at -0.548. The genetic predictor to its left (`gp_additive['plc']`) has a value of 1, and to its right (`gp_additive['wg789a']`) a value of -1. As the marker on the right is much closer (5.4cM away) than the one of the left (18.7cM away) the estimate of `gp_additive['glx']` is shifted from 0, the midpoint, towards to -1, the value of the closer neighbouring genetic predictor.

Row	_UnitLabels_	gp additive['plc']	gp additive['glx']	gp additive['wg789a']	
12	dh012	-1	-1	-1	
13	dh013	-1	-0.980252	-1	
14	dh014	1	-0.548174	-1	
15	dh015	-1	1	-1	

Figure 6.4: Excerpt from the spreadsheet of additive genetic predictors constructed from the Steptoe-Morex marker data.

Genetic predictors can be calculated for `F2`, double haploid (`DH1`), back-cross (`BC1`), recombinant inbred (`RILn`), back-cross inbred (`BCxSY`) and cross-pollinating (`CP`) populations. For a `RILn` population, the number of generations is supplied using the Number of generations: option in the Calculate Genetic Predictors menu (Figure 6.1). For a `BCxSy` population, the number of back-crosses and selfings is supplied using the Number of backcrosses: and Number of selfings: options, respectively.

6.1.2 Models for detecting QTLs

After calculating the genetic predictors, we can start building the QTL model. Essentially this involves finding a set of genetic predictors that best describe the phenotypic variation. The search for QTLs is done by testing the presence of a QTL at different positions on the chromosomes (as defined by the maximum step size along the genome).

Several linear regression methods have been proposed for QTL detection (e.g. Lander and Botstein, 1989; Haley and Knott, 1992), and these form the basis of the procedures in the Genstat QTL system. For a review see Collard *et al.* (2005). In Genstat, models for QTL detection are implemented in the mixed model framework with each genetic predictor (a potential QTL) tested as a fixed effect. There are three putative QTL detection methods available in the menu. In increasing complexity they are: marker regression, simple interval mapping (SIM) and composite interval mapping (CIM). We describe these in turn.

QTL analysis in Genstat requires a single phenotypic value for each genotype (or line) in each environment. These values will usually be predicted trait means obtained from a preliminary analysis of each trial (as described in Chapter 3).

Trait means from the Steptoe-Morex trial are held in file `SxM_pheno.csv`. We'll use the quantitative trait `yield` to illustrate QTL analysis in Genstat. Import the phenotypic data set following Section 2.1.1.2. Before QTL analysis, the quality of the phenotypic data should be assessed using the Data Exploration menu (see Section 2.4.1). Of particular interest is the amount of phenotypic variation in the population and the presence of any suspicious observations (i.e. outliers). For this example, there are no missing values for `yield`. The mean `yield` is 6.9 ton/ha with phenotypic variance 0.48 $(\text{ton/ha})^2$. A histogram and boxplot of `yield` indicate the data are Normally distributed and doesn't contain any extreme outliers.

6.1.2.1 Marker regression

In marker regression (also known as marker based QTL detection), phenotype is regressed on the genetic predictor separately for each marker position. The model for a single trait, single environment data set can be expressed as:

$$y_i = \mu + \alpha x_i + \varepsilon_i + e_i \qquad \text{Equation 1}$$

where

y_i is the trait mean for genotype i

μ is the overall mean

α is the QTL effect at the marker position being tested

x_i is the genetic predictor for genotype i at the marker position being tested

ε_i is the genetic residual for genotype i (or residual if unit errors are omitted)

e_i is the unit error for genotype i.

The model is fitted using REML, with `Genotype` fitted as random (representing ε_i) and the QTL effect fitted as fixed. The ε_i are assumed to follow a Normal distribution with mean 0 and variance σ^2.

The unit error, e_i, represents the uncertainty on the trait mean (see Section 3.2). However, if the unit error is unknown, e_i and the *genetic residual*, ε_i, cannot be separately estimated. We therefore omit the term e_i from the model, which means ε_i now represents the *residual* for genotype i.

Equation 1 is readily extended to include dominance effects (for `F2` and `CP` populations) and additive effects for the 2nd parent (for `CP` populations). See Section 6.2.

At each marker position the model is fitted and the QTL effect (α) tested using a Wald test (Section 10.6; Searle *et al.*, 1992; Verbeke and Molenberghs, 2000). The associated probability value (on the -$\log_{10}$ scale) is plotted against the position on the chromosome to produce a profile plot (also known as a Manhattan plot) used for interpretation.

The Single Trait Linkage Analysis (Single Environment) window (Figure 6.5) can be accessed by either:

- Stats | QTLs (Linkage/Association) | QTL Analysis | Single Trait Linkage Analysis (Single Environment); or,
- in the QTL Data View via the shortcut QTL analysis | Single Trait Linkage Analysis (Single Environment).

Genstat will automatically populate the input fields of this window using data from the QTL Data Space.
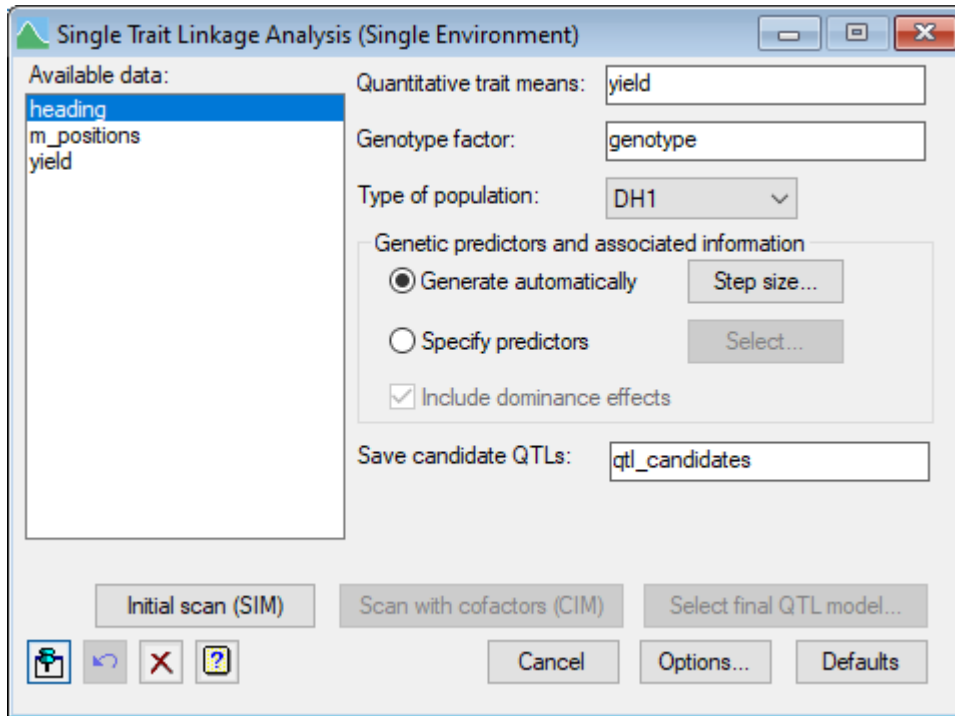
Figure 6.5: Menu for performing a single trait, single environment QTL analysis on the quantitative trait `yield` from the double haploid Steptoe-Morex population.

To perform marker regression, we use only genetic predictors calculated at marker positions. This is achieved from the Single Trait Linkage Analysis (Single Environment) window either by:

- setting the Step size to 10^6 (the default); or,
- using Specify predictor to select the genetic predictors calculated at markers positions in Section 6.1.1, `gp_additive` (Figure 6.6). The fields in the Genetic predictors and associated Information window are automatically filled using data from the QTL Data Space. As the Steptoe-Morex population is double haploid, the fields for Additive effects 2nd parent: and Dominance effects: are unavailable.
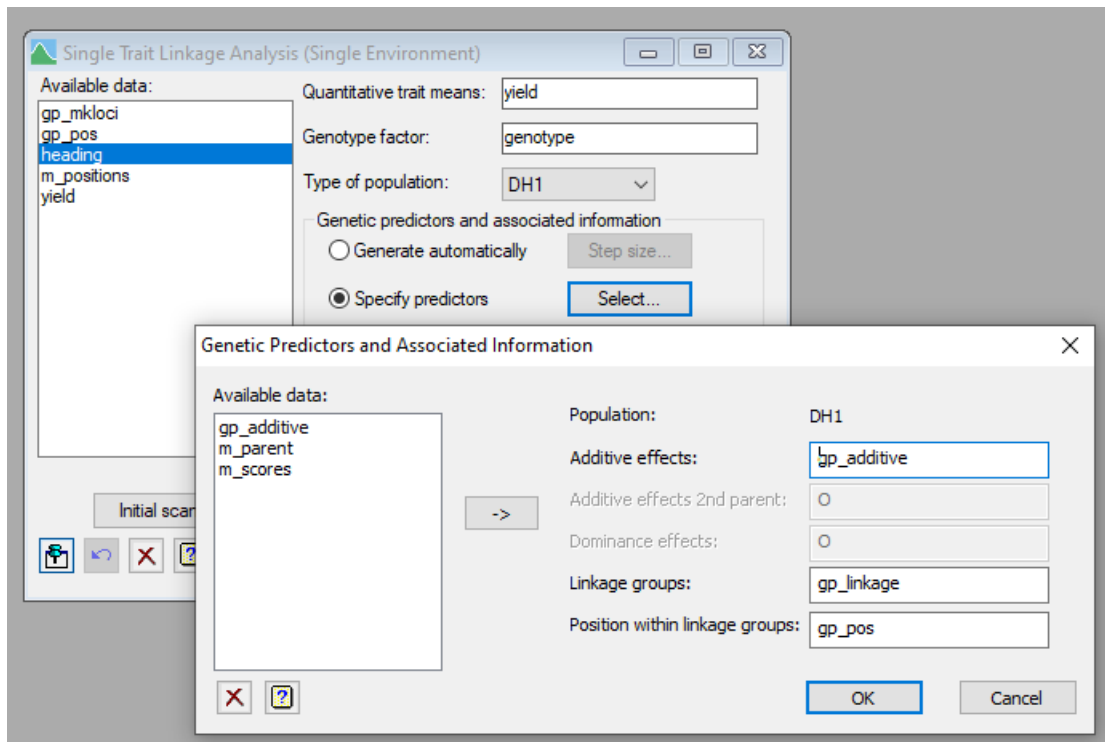
Figure 6.6: Window activated by Specify predictors, used to supply genetic predictors and associated information for a QTL analysis.

Settings for the QTL analysis can be modified by clicking the Options button (see Figure 6.7). The default options are to display a summary of the QTLs retained in the model, produce monitoring output detailing the progress of the analysis, and to produce a profile plot from the Wald tests of the QTL effects.

Genstat's QTL detection methods necessitate conducting multiple tests along the genome. The Threshold box specifies the method used to adjust for multiple comparisons (for details, see Section 6.1.4). By default the Li and Ji (2005) method is used with an overall significance level of 0.05.

If available unit errors should be included in the analysis using the option Include unit errors: (see Section 3.2). The other options will be explained in more detail as they become relevant. Detailed information on all options can be accessed via the help icon (▣). Click OK to close the Linkage Analysis Options window.
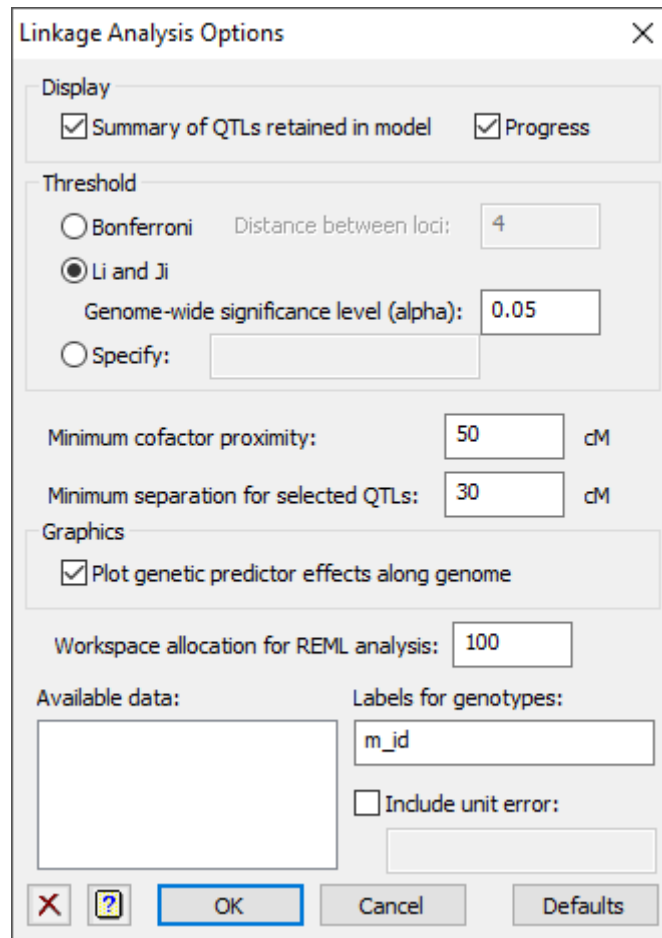
Figure 6.7: QTL Linkage Analysis Options window.

Marker regression can now be performed by clicking the Initial scan (SIM) button on the Single Trait Linkage Analysis (Single Environment) window (Figure 6.5). The output from this analysis includes a summary table containing the significant genetic predictors, the subset of these identified as QTL candidates, and a profile plot (Figure 6.8).

```
Summary
=======

Trait: yield
------------

The following loci have a test statistic larger than THRESHOLD=3.049

The test statistic is based on 1 set of predictors
```

```
        Locus       IdLocus Chromosome    Position   -Log10(P)
           18         chs1b          2        16.1        6.52
           19          rbcs          2        27.2        5.59
           20          abg2          2        41.2        6.52
           21        abg459          2        47.3        4.43
           37       abg703a          3        83.6        3.14
          105          rrn2          7        48.2        3.61
          106          ltp1          7        52.8        3.35
          107           ale          7        68.2        8.22
          108        abc302          7        78.2        6.93
          109        cdo57b          7        92.2        4.29


Selection of QTL candidates
===========================

The following candidates have been selected

     Locus    IdLocus  Chromosome     Position
        18      chs1b           2         16.1
        37    abg703a           3         83.6
       107        ale           7         68.2
```

The profile plot displays p-values from Wald tests (on the $-\log_{10}$ scale) of the QTL effects along the chromosomes (Figure 6.8). This is analogous to the logarithm of odds ratio (LOD) profile or Manhattan plot produced in other QTL software. The red horizontal line represents the threshold level, above which the null hypothesis of no QTL effect is rejected (see Section 6.1.4). Each chromosome (or linkage group) is depicted by a different colour. In the lower panel, the location of the QTL effects exceeding threshold are shown. The point is coloured according to which parent provided the allele for a high yield (i.e. the superior allele). The intensity of the colour represents the magnitude of the QTL effect.

Of the 116 genetic predictors (i.e. markers), 10 have an associated probability value (on the $-\log_{10}$ scale) larger than the threshold, in this case 3.049. To identify candidate QTLs, Genstat selects peaks in the profile plot (Figure 6.8) that meet the threshold and the minimum separation distance criteria set in the options (Figure 6.7). We've accepted the default, which sets Minimum separation for selected QTLs: to 30cM. If two peaks are closer together than 30cM then only the one with the highest peak will be taken forward as a candidate position. Of the 10 genetic predictors that meet threshold, only 3 are more than 30cM apart and therefore identified as candidate QTLs (chs1b, abg703a, ale). By default, candidate QTLs are saved to qtl_candidates (see Figure 6.5).
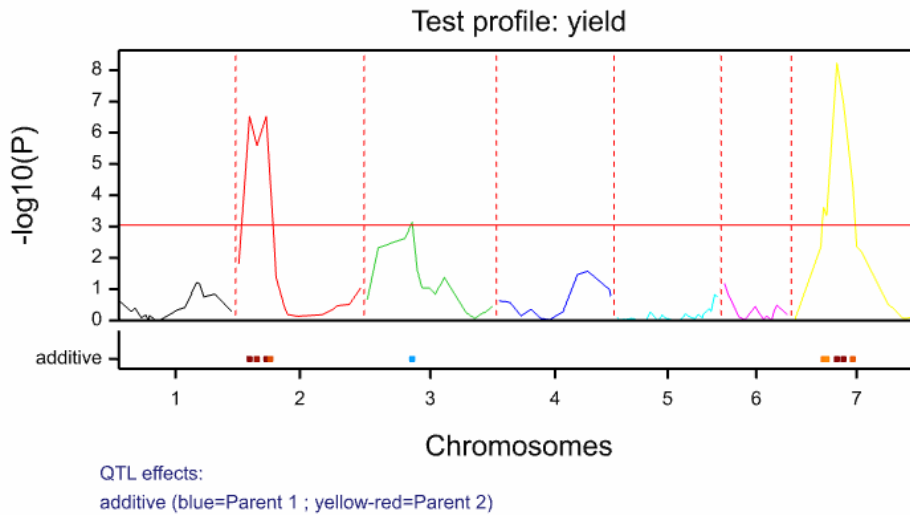
Figure 6.8: Profile from marker regression of `yield` for the Steptoe-Morex double haploid barley population.

Marker regression, although a simple technique useful for QTL identification, has two key limitations: 1) the residual term, ε_i, contains genetic variation due to QTLs segregating elsewhere in the genome and 2) the size of the QTL effect is confounded with its distance from the tested marker. Simple interval mapping (SIM) was proposed by Lander and Botstein (1989) to overcome confounding of the QTL effect with location.

6.1.2.2 Simple interval mapping

In marker regression QTLs are only assessed at marker locations. If the true position of a QTL is in-between markers, QTL effects will be underestimated and the most plausible location biased towards the marker positions. An alternative is to calculate genetic predictors in-between markers. These are known as "*pseudo-markers*". By using pseudo-markers, in addition to the real markers, better coverage of the genome is achieved, resulting in improved estimates of QTL locations and effects. This is especially true when the number of markers is not large.

In simple interval mapping (SIM) (Lander and Botstein, 1989), QTL effects are estimated as the regression coefficients on genetic predictors (α) calculated at intervals along the genetic map, as well as at marker positions (i.e. in Equation 1, x_i is extended to include locations between markers).

To perform SIM in Genstat, open the Stats | QTLs (Linkage/Association) | QTL Analysis | Single Trait Linkage Analysis (Single Environment) menu and click on the Step size button

(Figure 6.5). We'll set the step size to 10cM (Figure 6.9). This generates genetic predictors in-between markers, making sure that the gap between two consecutive genetic predictors is not larger than 10cM, in addition to at marker positions. Setting the step size to a small value will produce more detailed plots in the case of sparse maps, but will have little effect for dense maps. The smaller the step, the larger the number of evaluation points, and therefore the computation time required in the QTL mapping stage will be longer. If you do not change the step size from the default, 10^6, then marker regression will be performed. Click Initial scan (SIM) to perform the analysis.
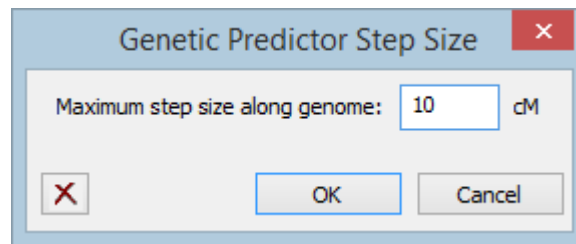
Figure 6.9: Changing the step size to calculate genetic predictors between marker positions for use in SIM.

Alternatively, genetic predictors calculated from the Calculate Genetic Predictors menu, where Options was used to specify the step size (see Section 6.1.1), can be inputted using Specify predictor (Figure 6.6).

The (re)calculated additive genetic predictors can be accessed from the QTL Data View (see Figure 6.2). The notation C*l*P*k* labels the genetic predictor calculated on linkage group (chromosome) l at position k. Using a maximum step size of 10cM, 174 genetic predictors are calculated; 116 from markers and 58 from pseudo-markers.

The output from a SIM analysis of `yield` is given below:

```
Summary
=======

Trait: yield
------------

The following loci have a test statistic larger than THRESHOLD=3.114

The test statistic is based on 1 set of predictors

       Locus     IdLocus Chromosome    Position   -Log10(P)
          26        C2P8          2         8.1        4.28
```

```
27        chs1b         2        16.1        6.52
28        C2P22         2        21.6        6.89
29         rbcs         2        27.2        5.59
30        C2P34         2        34.2        6.81
31         abg2         2        41.2        6.52
32        abg459        2        47.3        4.43
57        C3P57         3        56.9        3.12
61       abg703a        3        83.6        3.14
155        rrn2         7        48.2        3.61
156        ltp1         7        52.8        3.35
157       C7P61         7        60.5        6.14
158         ale         7        68.2        8.22
159       abc302        7        78.2        6.93
160        C7P85        7        85.2        6.41
161       cdo57b        7        92.2        4.29


Selection of QTL candidates
===========================


The following candidates have been selected

    Locus    IdLocus  Chromosome    Position
       28      C2P22           2        21.6
       61    abg703a           3        83.6
      158        ale           7        68.2
```
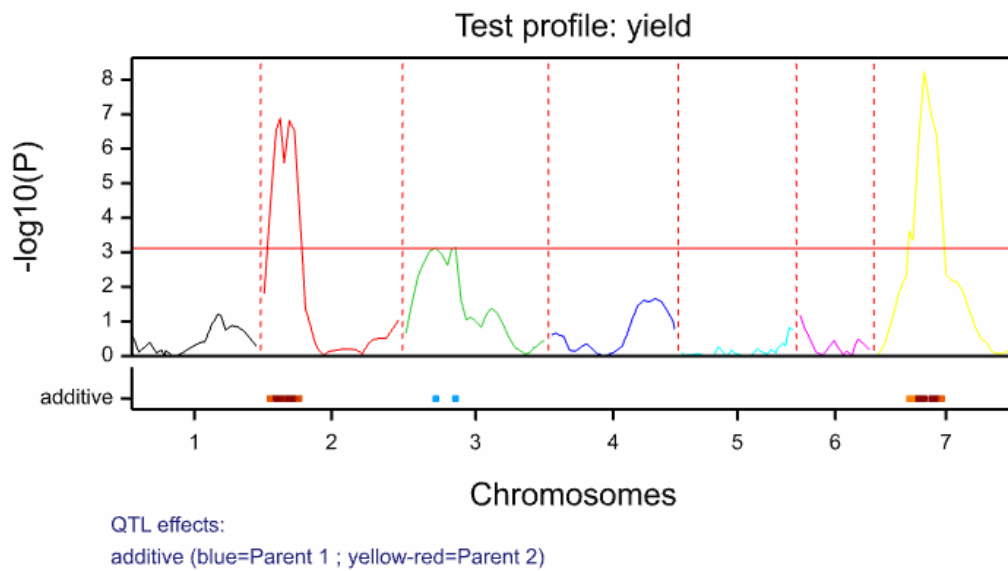

Figure 6.10: Profile plot from SIM of `yield`, with a step size of 10cM.

Of the 174 genetic predictors, 16 have an associated probability value (on the -$\log_{10}$ scale) larger than threshold, and 3 of these have been identified as candidate QTLs (`C2P22`, `abg703a` and `ale`). Both SIM and marker regression have identified candidate QTLs on chromosome 3 and 7 at marker loci. However, the estimated location of the candidate QTL on chromosome 2 has changed from marker `chs1b` with position 16.1 to pseudo-marker `C2P22` with position 21.6. The profile plot is given in Figure 6.10.

6.1.2.3 Composite interval mapping

In plant breeding we expect phenotypic trait variation to be the result of a large number of QTLs, each one with a minor effect. Although SIM removes the confounding of QTL effect with location, the residual term, ε_i, still contains genetic variation due to QTLs segregating elsewhere in the genome. Composite interval mapping (CIM) enhances the precision and power of QTL detection by introducing a number of genetic predictors, referred to as cofactors, to control for background genetic variation (i.e. the variation caused by QTLs outside the region where the QTL is being tested). CIM was simultaneously proposed by Zeng (1994) and Jansen and Stam (1994).

CIM is a multi-QTL model:

$$y_i = \mu + \sum_{c \in C} \alpha_c^* x_{ic} + \alpha x_i + \varepsilon_i + e_i \qquad \text{Equation 2}$$

where

C is the set genetic predictors used as cofactors, $c = 1, …, C$
y_i is the trait mean for genotype i
μ is the overall mean
α_c^* is the QTL effect for cofactor c
x_{ic} is the genetic predictor of cofactor c for genotype i.
α is the QTL effect at the position on the genome being tested
x_i is the genetic predictor for genotype i at the position being tested
ε_i is the genetic residual for genotype i (or residual if unit errors are omitted)
e_i is the unit error for genotype i.

As for marker regression and SIM, the model is fitted using REML with `Genotype` random (representing ε_i) and the QTL effects fitted as fixed. If estimates of unit error are not available, the term e_i is omitted from the model and the term ε_i represents the residual.

In Genstat we perform CIM, after an initial scan for QTLs by SIM (or marker regression), using candidate QTLs detected by SIM (or marker regression) as cofactors. CIM can then be repeated, using candidate QTLs detected by the previous CIM scan as

cofactors, until the list of candidate QTLs does not change. In general one or two rounds of CIM are usually sufficient.

To perform a CIM analysis in Genstat open the Stats | QTLs (Linkage/Association) | QTL Analysis | Single Trait Linkage Analysis (Single Environment) menu. The Scan with cofactor (CIM) button will be activated after an initial scan QTLs by SIM (or marker regression) (see Figure 6.5).

After performing a SIM scan on the Steptoe-Morex `yield` data, using a step size of 10cM (Section 6.1.2.2), click on Scan with cofactors (CIM) to open the Candidate QTL (cofactors) window (Figure 6.11).
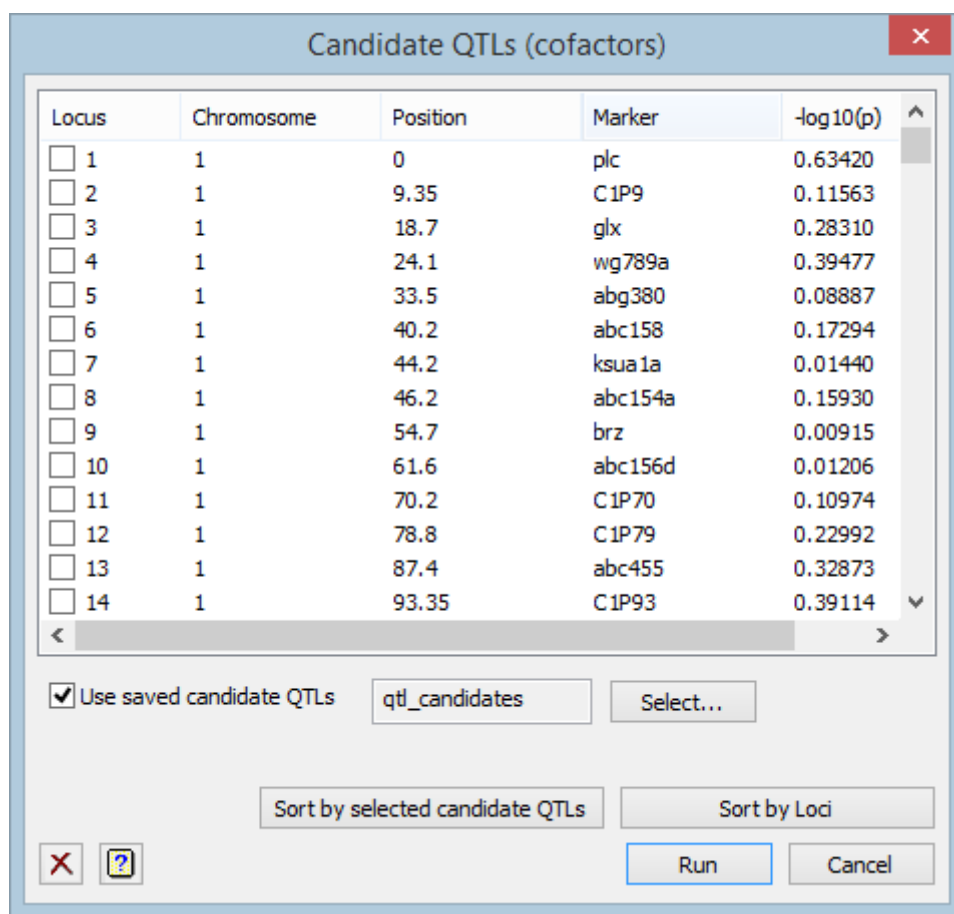
Figure 6.11: Candidate QTL (cofactors) window used to select cofactors to perform composite interval mapping (CIM).

The Candidate QTL (cofactors) window contains the complete list of genetic predictors (in this case 174) with their locus number, chromosome number, position, marker name

and -$\log_{10}$(p-value) from the last QTL scan (here, from SIM). This window is used to select cofactors. Automatically the candidate QTLs from the last QTL scan are selected (in our case `C2P22`, `abg703a` and `ale`) but you can modify this by removing or adding cofactors using the check box. To view the selected cofactors click on Sort by selected candidate QTLs. To re-order by loci click Sort by Loci. Leave the cofactors selected by default and run the analysis by clicking the Run button.

The output contains the profile plot (Figure 6.12), the list of genetic predictors than have a -$\log_{10}$(p-value) higher than threshold, and a list of candidate QTLs.

```
Summary
=======


Trait: yield
------------


The following loci have a test statistic larger than THRESHOLD=3.114

The test statistic is based on 1 set of predictors

        Locus      IdLocus Chromosome    Position   -Log10(P)
           26         C2P8          2         8.1        4.42
           27        chs1b          2        16.1        6.83
           28        C2P22          2        21.6        7.96
           29         rbcs          2        27.2        7.09
           30        C2P34          2        34.2        9.06
           31         abg2          2        41.2        8.95
          154       abg395          7        45.6        3.36
          155         rrn2          7        48.2        4.22
          156         ltp1          7        52.8        3.64
          157        C7P61          7        60.5        6.76
          158          ale          7        68.2        9.13
          159       abc302          7        78.2        8.24
          160        C7P85          7        85.2        6.73
          161       cdo57b          7        92.2        3.87


Selection of QTL candidates
===========================


The following candidates have been selected

     Locus    IdLocus   Chromosome      Position
        30      C2P34            2          34.2
       158        ale            7          68.2
```
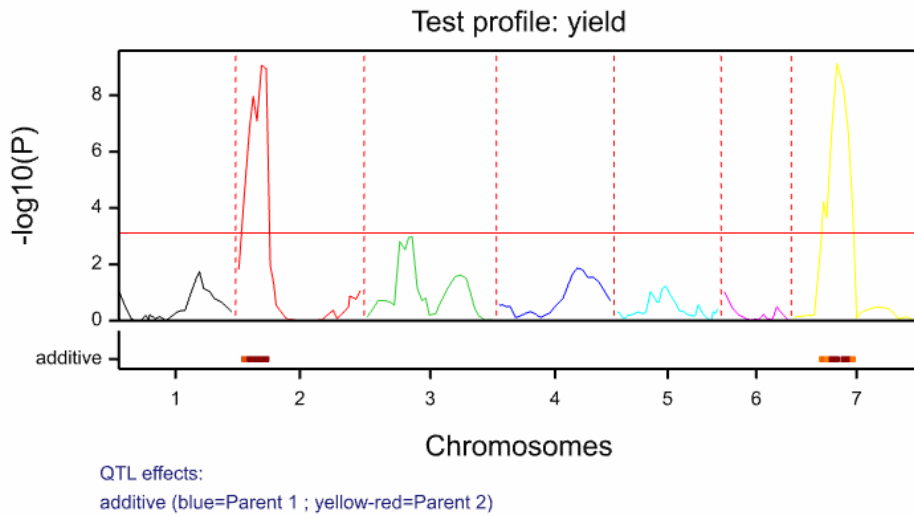
Figure 6.12: Profile plot from one round of CIM on yield, using the candidate QTLs from SIM as cofactors.

After one round of CIM, candidate QTLs are detected on chromosomes 2 and 7. However, the position of the candidate QTL on chromosome 2 has changed from 26.1 to 34.2cM. Also note, the small effect detected by SIM on chromosome 3 is no longer significant. We repeat CIM until we don't see any more changes in the list of candidate QTLs. On the second round, marker `abg703a` is reinstated as a candidate. On round three, no further changes are made. After CIM, the list of candidate QTLs to be considered in the final QTL model are:

```
Selection of QTL candidates
===========================

The following candidates have been selected

    Locus   IdLocus  Chromosome    Position
       30     C2P34           2        34.2
       61    abg703a           3        83.6
      158       ale           7        68.2
```

When performing a CIM analysis it is important to avoid co-linearity between a potential QTL and a cofactor. This is done by temporarily removing cofactors from the model when testing a QTL in a nearby region. By default, Genstat specifies a 50cM region around the potential QTL. This can be changed in the Linkage Analysis Options window using the Minimum cofactor proximity: setting (see Figure 6.7). The sensitivity of the CIM

results to this setting should be tested. When Minimum cofactor proximity: is set to a large value, i.e. 10^6, all cofactors on the same chromosome as the potential QTL are excluded, a strategy known as restricted CIM.

6.1.2.4 Final QTL model

After scanning the genome to detect candidate QTLs (by marker regression, SIM, or preferably CIM), the final step in QTL mapping consists of fitting a final multi-QTL model. Here phenotypic variation is modelled as the result of the contributions from several QTLs using multiple linear regression. By default the final model is selected using backward selection; here all candidate QTLs are fitted in the same model and then tested one-by-one to determine if their contribution to the model is significant.

We can write the final multi-QTL model as:

$$y_i = \mu + \sum_{q \in Q} \alpha_q x_{iq} + \varepsilon_i + e_i \qquad \text{Equation 3}$$

where

Q is the set of QTLs, $q = 1, \dots, Q$
y_i is the trait mean for genotype i
μ is the overall mean
α_q is the effect of QTL q
x_{iq} is the genetic predictor of QTL q for genotype i
ε_i is the genetic residual for genotype i (or residual if unit errors are omitted), assumed to follow a Normal distribution with mean 0 and variance σ^2
e_i is the unit error for genotype i.

The model is fitted using REML, with `Genotype` random (representing ε_i) and the QTL effects (α_q, $q = 1, \dots, Q$) fitted as fixed. If estimates of unit error are not available, the e_i term is omitted from the model and the term ε_i represents the residual.

To perform the analysis click on Select final QTL model in the Single Trait Linkage Analysis (Single Environment) window (Figure 6.5). This will open the Select Final QTL Model window where you can select what output to display, specify whether to use backward selection to determine the final QTL model, and save results from the final QTL model (Figure 6.13).

To view (or modify) the candidate QTLs tested in the final model, click on Candidate QTLs. We use the three candidate QTLs detected by CIM (`C2P34`, `abg703a`, `ale`). Click Run to perform the analysis.
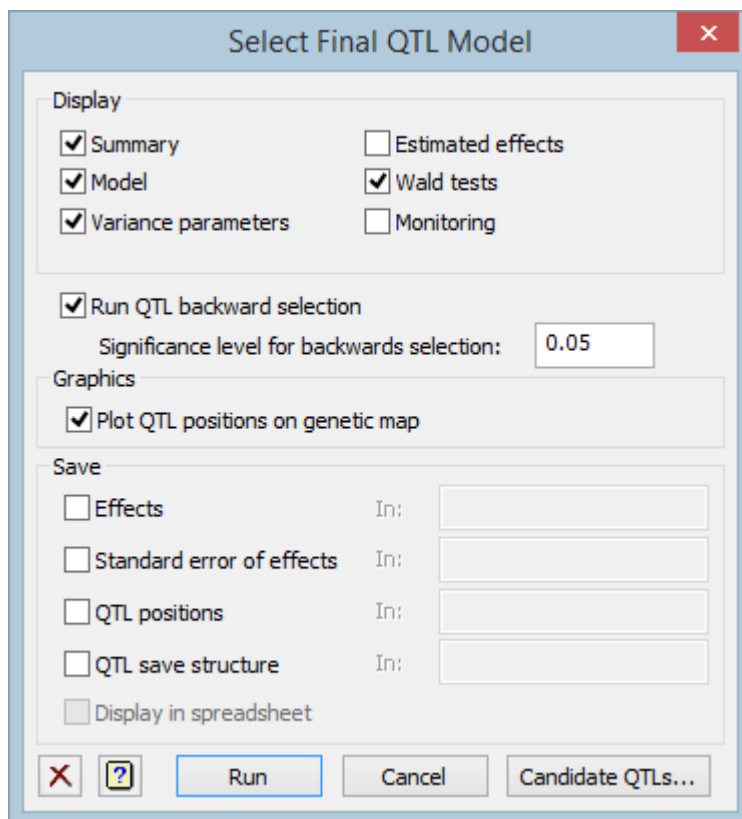
Figure 6.13: Window for fitting a final QTL model to estimate QTL effects in single environment trials.

The output for the final multi-QTL model is:

```
Estimation of QTL effects from a single-environment trial
=========================================================



REML variance components analysis
=================================

Response variate:  yield
Fixed model:       Constant + QTL[30] + QTL[61] + QTL[158]
Random model:      genotype
Number of units:   150

genotype used as residual term
```

```
Sparse algorithm with AI optimisation


Residual variance model
-----------------------

Term                      Model(order)  Parameter      Estimate      s.e.
genotype                  Identity      Sigma2            0.297    0.0348


Tests for fixed effects
-----------------------

Sequentially adding terms to fixed model

Fixed term               Wald statistic  n.d.f.   F statistic  d.d.f.    F pr
QTL[30]                           37.94       1         37.94   146.0  <0.001
QTL[61]                           17.67       1         17.67   146.0  <0.001
QTL[158]                          40.30       1         40.30   146.0  <0.001

Dropping individual terms from full fixed model

Fixed term               Wald statistic  n.d.f.   F statistic  d.d.f.    F pr
QTL[30]                           37.61       1         37.61   146.0  <0.001
QTL[61]                           13.61       1         13.61   146.0  <0.001
QTL[158]                          40.30       1         40.30   146.0  <0.001

* MESSAGE: denominator degrees of freedom for approximate F-tests are calculated
using  algebraic  derivatives  ignoring  fixed/boundary/singular  variance
parameters.


Summary
=======


Trait: yield
Population type: DH1
Number of genotypes: 150
Number of linkage groups: 7
Number of markers: 116

List of QTLs
============

 Locus  Locus           Linkage Position -log10(P)
   no.   name             group
    30  C2P34               2    34.20     8.114
    61  abg703a             3    83.60     3.499
   158  ale                 7    68.20     8.586
```

```
QTL effects
===========

 Locus  Locus              %Expl.      Add. High value  s.e.
   no.  name                 Var.      eff.     allele
    30  C2P34              17.662     0.292    Steptoe 0.048
    61  abg703a             5.894     0.169      Morex 0.046
   158  ale                17.019     0.287    Steptoe 0.045


Estimated lower and upper bounds of QTL positions
=================================================

 Locus  Locus             Lower Position    Upper
   no.  name              bound             bound
    30  C2P34             0.000   34.200  183.100
    61  abg703a          16.300   83.600  205.000
   158  ale               4.700   68.200  191.200
```

The final QTL model fits three QTLs, all of which are highly significant (p<0.001; from dropping the individual QTL terms from the full fixed model). The estimated QTL effects are given in the table `QTL effects`, with the column `High value allele` indicating which parent is providing the high yielding allele for each QTL. For example, the estimated effect for the QTL `C2P34`, is 0.292 ton/ha (Steptoe) with standard error 0.048 ton/ha. Therefore replacing the Morex parental allele by the Steptoe parental allele at QTL `C2P34` is expected to result in an increase yield of 0.292 ton/ha (the expected difference in yield between the two homozygous types is twice as large). The final table in the output gives the estimated locations of the QTLs with 95% confidence intervals.

In Figure 6.14 the significant QTLs are plotted on a genetic map.
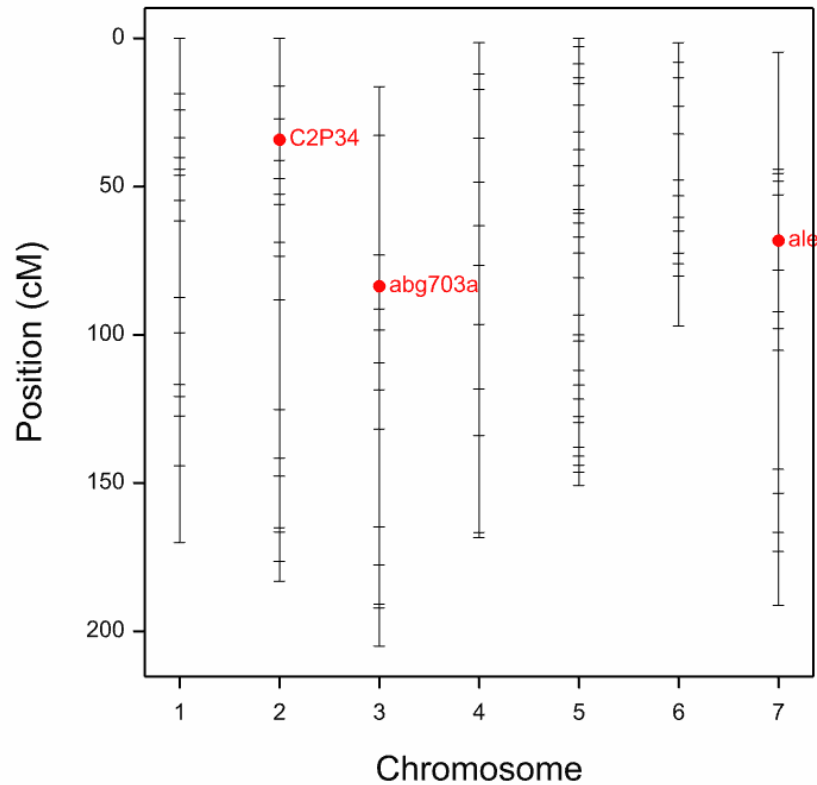
Figure 6.14: Genetic map of the Steptoe-Morex data with detected QTLs shown.

6.1.3   Accounting for uncertainty in trait means

Linkage analysis in Genstat is performed on trait means. However, if the genotypes are unequally replicated, incorporating the unit error, a measure of precision of the trait means (see Section 3.2), in QTL analysis can improve detection. Unit errors are included in a QTL analysis via the Include unit errors: field in the Linkage Analysis Options window (Figure 6.7). If a unit error data structure has been stored in the QTL Data Space then the name of that structure will be automatically entered into the input field.

6.1.4   Multiple comparisons

Genstat's QTL models are implemented in the mixed model framework with each genetic predictor (potential QTL) tested as a fixed effect. The systematic genome-wide testing of each genetic predictor in marker regression, SIM, and CIM introduces the problem of multiple testing. That is, the genome-wide Type I error rate is inflated and the probability of declaring a non-significant QTL as significant is increased. The methods implemented in Genstat to moderate the false discovery rate are a Bonferroni correction and a method proposed by Li and Ji (2005). The Bonferroni correction adjusts the genome-wide error

rate for the number of tests performed. However, it assumes (incorrectly) that independent tests occur at a fixed distance on the genome (default 4cM). This results in too conservative a threshold for QTL detection. Genstat's default is to use the method proposed by Li and Ji (2005); a Bonferroni correction based on the effective number of independent tests.

Both methods result in a genome-wide significance threshold, expressed as a p-value on the -$\log_{10}$ scale. This determines the critical value to reject the null hypothesis of no QTL effect. The genome-wide Type I error rate, which defaults to 0.05, can be specified in the Genstat Linkage Analysis Options window (Figure 6.7). The Linkage Analysis Options window also allows you to specify your own threshold value, expressed as a p-value on the -$\log_{10}$ scale.

6.2 Dominance and additive effect of the second parent

The models for detecting QTLs described in Section 6.1 include only additive genetic predictors. However, in the case of F2 or cross-pollinating (CP) populations, dominance genetic predictors can also be modelled. The equations for the statistical models presented in Section 6.1 can be easily modified to accommodate dominance effects. For example, to incorporate dominance effects into the marker regression (or SIM) model, we modify Equation 1 to:

$$y_i = \mu + \alpha^a x_i^a + \alpha^d x_i^d + \varepsilon_i + e_i \qquad \text{Equation 4}$$

where

α^a is the additive QTL effect at the position being tested

x_i^a is the additive genetic predictor for genotype *i* at the position being tested

α^d is the dominance QTL effect at the position being tested

x_i^d is the dominance genetic predictor for genotype *i* at the position being tested.

Including dominance effects into QTL detection models is straightforward - simply check the Include dominance effects: on Single Trait Linkage Analysis (Single Environment) window (Figure 6.5).

Additive genetic predictors of the second parent are also readily incorporated when conducting a QTL analysis on a cross-pollinating (CP) population (see Chapter 8).

When dominance and/or additive effects of the second parent are modelled, the term "*QTL effect*" often refers to the combined effect of all genetic predictors modelled at that locus: i.e. the additive, dominance, and additive effects of the second parent.

6.3   References

Collard, B.C.Y, Jahufer, M.Z.Z, Brouwer, J.B., & Pang, E.C.K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, **142**,169–196.

Haley, C.S., & Knott, S.A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315-324.

Jansen, R. C., & Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, **136**, 1447-1455.

Jiang, C.J., & Zeng, Z.B. (1997). Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica*, **101**, 47-58.

Lander, E.S., & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185-199.

Li, J., & Ji, L. (2005) Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, **95,** 221-227.

Searle, S.R., Casella, G., & McCulloch, C.E. (1992). Variance components. Wiley, New York.

Verbeke, G., & Molenberghs, G. (2000). Linear mixed models for longitudinal data. Springer-Verlag Inc.: Berlin; New York.

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457-1468.

7 Linkage analysis: inbred population with multiple traits evaluated or multiple trials

In Chapter 6 linkage analysis, a method for detecting quantitative trait loci (QTLs), was illustrated for an inbred population with a single trait evaluated at a single site. However, linkage analysis can also be performed across multiple environments, with estimation of $QTL \times$ environment (QTL×E) interactions, and for multiple traits in a single environment, allowing QTLs to be simultaneously tested for on multiple traits. Chapter 7 extends the analysis for inbred populations to these cases. For both data types, we describe the underlying statistical models and demonstrate QTL linkage analysis using Genstat.

In this chapter you will learn how to:

- detect QTLs in a single trait, multiple environment data set (Section 7.2)
- detect QTLs in a multiple trait, single environment data set (Section 7.3)

7.1   QTL linkage analysis in Genstat

Chapter 6 illustrated QTL analysis in Genstat for the simplest type of data set: single trait - single environment. In this chapter we extend QTL analysis to data sets where a single trait has been evaluated in multiple environments (Section 7.2); or where multiple traits have been assessed in a single environment (Section 7.3). For both types of data we i) describe the statistical theory underpinning Genstat's QTL linkage models, and ii) demonstrate QTL analysis in Genstat.

The analysis strategy used to detect and estimate QTL effects in these more complex data sets is very similar to that presented in Chapter 6. Namely:

1) predicted trait means are obtained from preliminary analyses of each trial (see Chapter 3)
2) genetic predictors are calculated for use as explanatory variables in the QTL models (see Section 6.1.1)
3) an initial genome-wide scan by marker regression or SIM (see Sections 6.1.2.1 and 6.1.2.2, respectively) is performed to obtain candidate QTL positions for use as cofactors in subsequent CIM scans
4) one or more rounds of CIM is performed, in which co-factors correct for QTLs that segregate elsewhere in the genome (see Section 6.1.2.3)
5) the final multi-QTL model is selected and fitted (see Section 6.1.2.4)

The key differences are, for multiple environment (single trait) data sets the scans test for environment-specific QTL effects, allowing for QTL by environment interactions (QTL×E). For multiple trait (single environment) data sets, the scans test for QTL effects on each trait simultaneously.

7.2   Single trait multiple environments

The statistical models for QTL detection for a single trait - single environment data set are described in Section 6.1.2. Here, we extend these models to multiple environment (single trait) data sets. For simplicity, we consider only additive genetic effects, although the models are easily modified to incorporate both dominance and additive effects of the second parent (refer to Section 6.2).

For multiple environments, the model for SIM (or marker regression) can be expressed as (refer to Section 6.1.2.2):

$$y_{ij} = \mu + E_j + \alpha_j x_i + \varepsilon_{ij} + e_{ij} \qquad \text{Equation 1}$$

where

y_{ij} is the trait mean for genotype i in environment j ($j = 1, \ldots, m$)

μ is the overall mean

E_j is the environment j main effect

α_j is the QTL effect for environment j at the position on the genome being tested

x_i is the genetic predictor for genotype i at the position being tested

ε_{ij} is the residual for genotype i in environment j

e_{ij} is the unit error for genotype i in environment j.

The model is fitted using REML, with the QTL ($\alpha_j, j = 1, \ldots, m$) and environment ($E_j$, $j = 1, \ldots, m$) effects fitted as fixed and `Genotype` fitted as a random (allowing specification of the variance-covariance matrix).

The residuals, ε_{ij}, representing the unexplained genotype and environment effects, are assumed be Normally distributed with mean 0 and variance-covariance structure VCOV(ε_{ij}). The matrix VCOV can either be modelled explicitly (i.e. with an unstructured model) or using the best variance-covariance model selected during a G×E analysis, as described in Chapter 4. The unit errors, e_{ij}, represent the uncertainty on the trait means (see Section 3.2). However, if estimates of the unit errors are not available, e_{ij} and ε_{ij} cannot be separately estimated (see Section 4.3).

The CIM model is formed by including cofactors into Equation 1 (refer to Section 6.1.2.3):

$$y_{ij} = \mu + E_j + \sum_{c \in C} \alpha_{jc}^* x_{ic} + \alpha_j x_i + \varepsilon_{ij} + e_{ij} \qquad \text{Equation 2}$$

where

C is the set genetic predictors used as cofactors, $c = 1, \ldots, C$

α_{jc}^* is the QTL effect for cofactor c in environment j

x_{ic} is the genetic predictor of cofactor c for genotype i.

We can write the final multi-QTL model as (refer to Section 6.1.2.4):

$$y_{ij} = \mu + E_j + \sum_{q \in Q} \alpha_{jq} x_{iq} + \varepsilon_{ij} + e_{ij} \qquad \text{Equation 3}$$

where

Q is the set of QTLs, $q = 1, \ldots, Q$

α_{jq} is the effect of QTL q in environment j.

x_{iq} is the genetic predictor of QTL q for genotype i.

We illustrate QTL analysis for single trait - multiple environment data sets using mean maize yields (`yld`) from the 8 environment CIMMYT maize trials held in file `F2maize_pheno.csv` (described in Section 1.3.2). Marker and map information for this F2 population are held in Flapjack files `F2maize_geno.txt` and `F2maize_map.txt`, respectively. Import and check the phenotypic and genotypic data (see Chapter 2). This F2 population compromises 211 individuals (genotypes) genotyped with 122 markers. The 122 markers have been mapped to 10 linkage groups ranging in length from 109cM to 266cM.

To resolve the discrepancy in the ordering of genotypes between the phenotypic and marker data sets run a compatibility check: Stats | QTLs (Linkage/Association) | Data Manipulation | Compatibility Check (see Section 2.3.1). Here, the new data structures have been suffixed by `_new` (Figure 7.1).
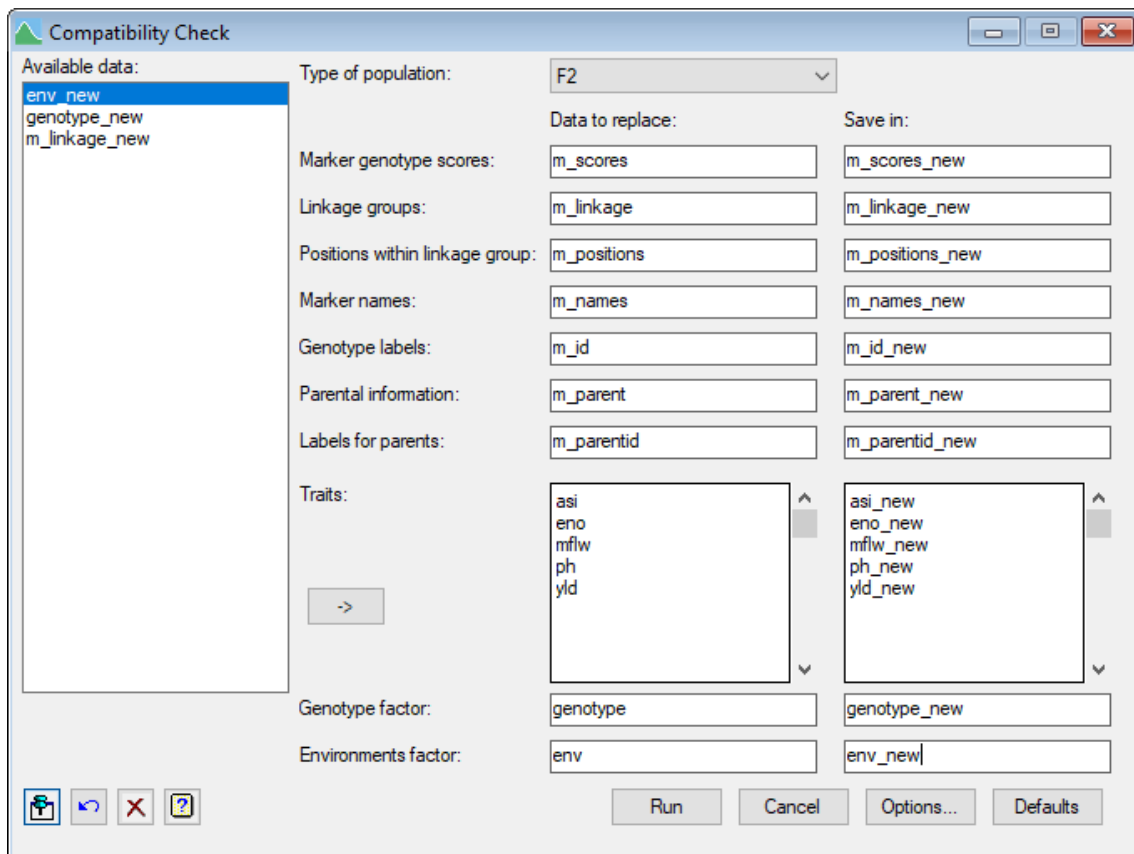
Figure 7.1: Performing a compatibility check on the phenotypic and genotypic CIMMYT maize data to resolved the discrepancy in the ordering of genotypes.

The Single Trait Linkage Analysis (Multiple Environments) window (Figure 7.2) can be accessed by either:

- Stats | QTLs (Linkage/Association) | QTL Analysis | Single Trait Linkage Analysis (Multiple Environments); or,

- in the QTL Data View via the shortcut QTL analysis | Single Trait Linkage Analysis (Multiple Environments).

Genstat will automatically populate the input fields using data from the QTL Data Space. The menu is similar to that of a single trait - single environment analysis (see Figure 6.5) except for the addition of the Environment factor: and Variance-covariance model: fields. In this illustration, we include only additive effects in the QTL models, however as this is an F2 population you could opt to include dominance effects (see Section 6.2). If unit errors are available, they should be included in the analysis by checking the Include unit errors: box accessed via the Options button (refer to Figure 6.7). Refer to Chapter 6 for further details on the input fields and Options.
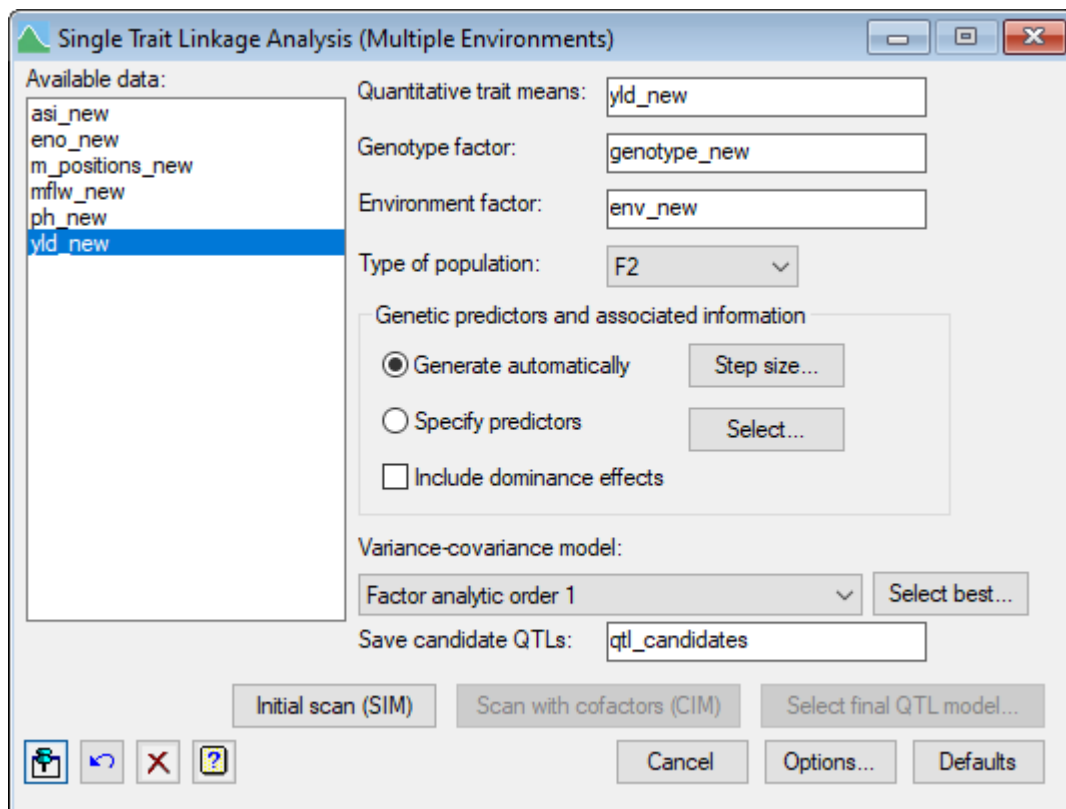
Figure 7.2: Menu for performing a single trait - multiple environment QTL analysis on yield (`yld_new`) from the CIMMYT maize trial data. For this population, `FA` (Factor analytic, order 1) is chosen as the best variance-covariance model (see Section 4.2).

The Variance-covariance model: field allows you to select the variance-covariance structure for modelling the variation between genotypes both across and within environments (see Section 4.1.3). If you've previously performed a genotype by environment analysis (described in Chapter 4), then the best model is selected by default. If not, the Select best button opens the Select Best Variance-covariance Model window (see Figure 4.3), enabling the selection of the best variance-covariance model based on either Schwarz Information Criterion (SIC) or Akaike Information Criterion (AIC). In this example, SIC identifies FA (Factor analytic, order 1) as the "*best model*" for yld_new (SIC = 17524). For full details on selecting the best variance-covariance model refer to Section 4.2.

To perform an initial scan by SIM with a maximum gap of 10cM between two consecutive genetic predictors (for example), use the Step size button to set the interval to 10 (see Figure 6.9). If you do not change the step size from the default, 10^6, then marker regression will be performed (see Section 6.1.2.1). We accept the default Options (see Figure 6.7), including the Li and Ji (2005) method to adjust the genome-wide significance level (default = 0.05) for multiple comparisons. Click Initial scan (SIM) to run the analysis.

Results from SIM, given below, list 26 positions where the associated $-\log_{10}$(p-value) is above threshold; here 3.21, calculated according to the Li and Ji method. The notation C*l*P*k* is used to denote position *k* on linkage group (chromosome) *l*. SIM has identified seven candidate QTLs (L085, L028, C2P36, C3P46, L071, L043 and C10P60). Their selection is according to the criterion specified in the options (in this example, a minimum distance of 30cM apart).

```
Summary
=======

Trait: yld_new
The following loci have a test statistic larger than THRESHOLD=3.21

    Locus IdLocus Chromosome    Position -Log10(P)
       14  C1P102          1       101.9      4.75
       15  C1P110          1       110.4      6.84
       16  C1P119          1       119.0      8.53
       17    L065          1       127.5      9.24
       18  C1P134          1       134.2     12.44
       19    L085          1       141.0     13.41
       20    L039          1       150.2      8.45
       21  C1P158          1       158.4      7.71
       22  C1P167          1       166.6      5.31
       35    L028          1       252.0      3.39
```

```
 42    C2P36          2          35.9         3.48
 70    C3P38          3          37.9         3.60
 71    C3P46          3          45.5         3.88
 72     L020          3          53.2         3.57
 73     L035          3          55.7         3.74
 74    C3P61          3          61.5         3.47
112     L071          4         136.6         3.42
113     L080          4         145.0         3.31
158   C6P117          6         117.2         3.35
159     L043          6         125.0         4.08
160   C6P133          6         133.4         3.33
236     L114         10          53.2         6.41
237    C10P60        10          60.1         8.44
238     L089         10          67.1         7.80
239    C10P76        10          76.2         7.09
240    C10P85        10          85.2         4.47


Selection of QTL candidates
===========================


The following candidates have been selected

    Locus    IdLocus  Chromosome     Position
       19      L085            1        141.0
       35      L028            1        252.0
       42     C2P36            2         35.9
       71     C3P46            3         45.5
      112      L071            4        136.6
      159      L043            6        125.0
      237    C10P60           10         60.1
```

The SIM profile plot, Figure 7.3, consists of two frames. The upper frame profiles the p-values (on the -log$_{10}$ scale) against linkage group position. A red horizontal line indicates the significance threshold. In the lower panel, the additive QTL effects are shown. The 8 environments are listed on the Y-axis, with coloured points aligning to positions with a test statistic exceeding threshold. The point is coloured according to which parent provided the allele for high `yld_new` (i.e. the superior allele); blue denotes that the high value allele originates from parent 1 and red from parent 2. The intensity of the colour represents the magnitude of the effect.

In this example, the candidate QTL identified in the middle of chromosome 1 (`L085`) has a significant additive effect in all environments except `LN96a` and `LN96b`. The high value allele is provided by parent 2 (red), except in environment `HN96b` where the superior allele is that of parent 1 (blue). This reversal of performance is an example of

QTL×E cross-over interaction (see Section 4.1.1). From the intensities of the colours we deduce that the QTL effect is strongest in `SS92a` (dark red) and weakest in `SS94a`.
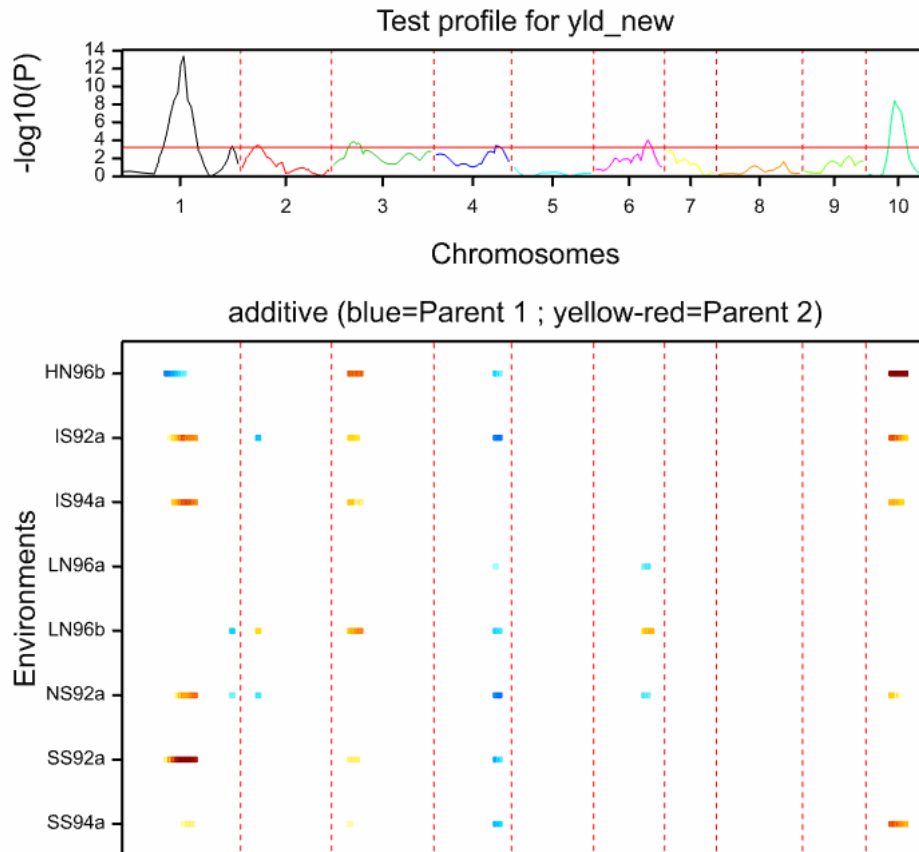
Figure 7.3: Profile plot from SIM of `yld_new`, with a step size of 10cM.

Note: failure to perform a compatibility check (Section 2.3.1) prior to QTL analysis results in the error message:

```
******** Fault 1, code UF 1, statement 99 in procedure QMQTLSCAN

Phenotypic and marker data are in a different order because labels of GENOTYPES
and IDMGENOTYPES are not identical.
```

After performing an initial genome-wide scan by SIM, the Scan with cofactors (CIM) and Select final QTL model buttons will activate. Before selecting the final multi-QTL model, a CIM scan is recommended. CIM performs a QTL search controlling for background genetic variation using cofactors (see Section 6.1.2.3), increasing the power to detect QTLs.

To scan for QTLs using CIM click on the Scan with cofactors (CIM) button. This opens the Candidate QTLs (cofactors) window (Figure 6.11). The candidate QTLs from the last QTL scan are automatically selected as cofactors, but any set of cofactors can be selected. Leave the cofactors selected by default and click Run (this could take some time). It is usual to repeat CIM until the list of candidate QTLs is unchanged (in this case, 1 round is sufficient).

The output produced is essentially the same as for SIM; a summary of results (now including the list of cofactors used) and a profile plot (Figure 7.4).

```
Summary
=======


Trait: yld_new
Specified cofactors are:
    Cofactor     IdLocus   Chromosome     Position
          19        L085            1       141.00
          35        L028            1       252.00
          42       C2P36            2        35.90
          71       C3P46            3        45.53
         112        L071            4       136.60
         159        L043            6       125.00
         237      C10P60           10        60.15


The following loci have a test statistic larger than THRESHOLD=3.21

    Locus IdLocus Chromosome    Position -Log10(P)
       16  C1P119          1       119.0      8.07
       17    L065          1       127.5      8.83
       18  C1P134          1       134.2     11.97
       19    L085          1       141.0     12.85
       20    L039          1       150.2      8.07
       21  C1P158          1       158.4      7.91
       41   C2P26          2        26.1      3.66
       42   C2P36          2        35.9      4.37
       43   C2P46          2        45.6      4.17
       44    L120          2        55.4      3.35
       69   C3P30          3        30.2      3.71
       70   C3P38          3        37.9      4.42
       71   C3P46          3        45.5      4.41
       72    L020          3        53.2      3.81
       73    L035          3        55.7      4.64
       74   C3P61          3        61.5      4.39
       75    L100          3        67.2      3.28
      112    L071          4       136.6      3.43
      159    L043          6       125.0      3.73
      236    L114         10        53.2      5.42
      237  C10P60         10        60.1      7.82
```

```
238    L089         10      67.1      7.74
239  C10P76         10      76.2      7.34


Selection of QTL candidates
===========================

The following candidates have been selected

    Locus    IdLocus  Chromosome      Position
       19       L085           1         141.0
       42       C2P36           2          35.9
       73       L035           3          55.7
      112       L071           4         136.6
      159       L043           6         125.0
      237      C10P60          10          60.1
```
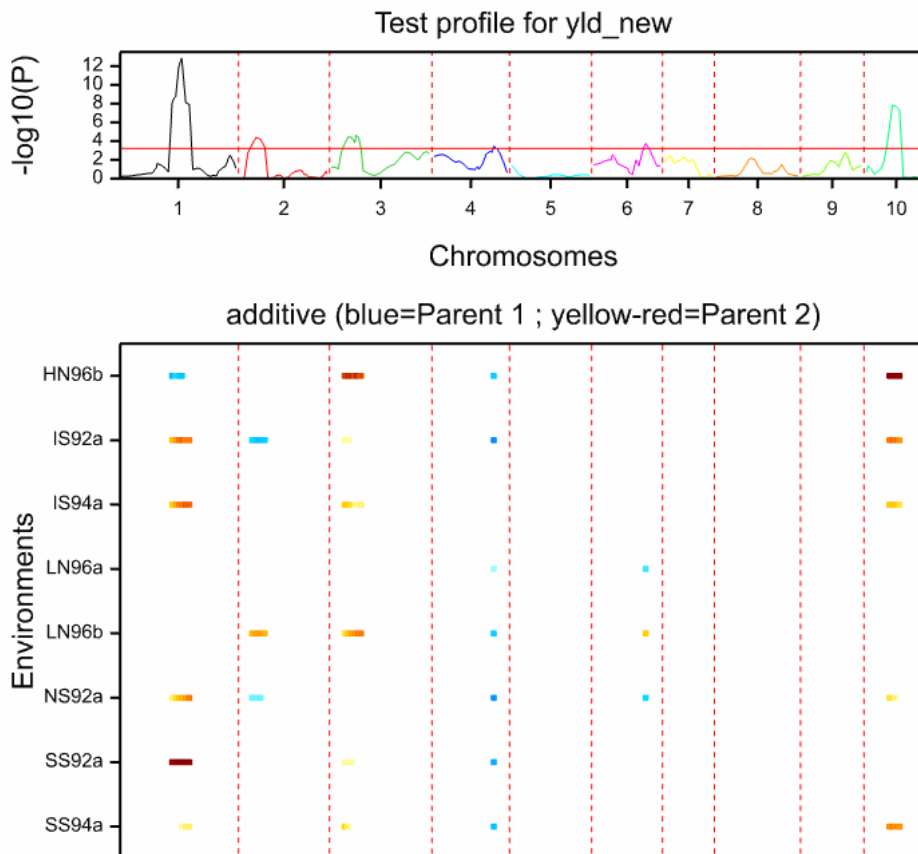

Figure 7.4: Profile plot from CIM of `yld_new`, after 1 round of SIM and 1 round of CIM.

Six candidate QTLs have been identified by CIM; one fewer than SIM with now only one candidate on chromosome 1. Furthermore, the position of the candidate QTL on chromosome 3 has altered slightly: from 45.5 to 55.7cM. The profile plot (Figure 7.4) indicates QTL×E cross-over interactions on chromosomes 1, 2, and 6.

Now that a set of candidate QTLs has been defined, the final multi-QTL model can be selected and fitted (see Section 6.1.2.4). Click on the Select final QTL model button to open the Select Final QTL Model window (Figure 7.5). From this window you can select what output to display, specify whether to use backward selection to determine the final model, set the variance-covariance model, and save results from the final QTL model. By default the variance-covariance model will be set to the model used for the QTL scan, and backward selection with significance level 0.05 will be used to select the final QTL model. Unselecting Run QTL backward selection will force all candidate QTLs to be retained in the final model and to have environment-specific effects. Click Run to select the final QTL model using backward selection.

Figure 7.5: Window for fitting a final multi-QTL model to a single trait - multiple environment data set.

The backward selection process compromises of two key steps:

1) **Determining which QTLs are significant in the multi-QTL model.** Starting with a model that contains all candidate QTLs, the procedure iteratively tests the importance of each candidate QTL, i.e. tests their effect conditional on the other QTLs in the model, leaving out those that are not significant. Here, "*QTL effect*" is the additive effect, combined, if modelled, with the additive effect of the second parent and/or the dominance effect. This process is repeated until all QTLs included in the model are significant.

2) **Testing for QTL×E interactions.** The procedure then tests whether the remaining QTLs exhibit significant interactions with environment, by partitioning the QTL effects into QTL main effects and QTL×E interactions. If the QTL×E interaction is not significant, then only the QTL main effect is retained in the model.

When dominance is fitted, the backward selection process has a third step:

3) **Testing for dominance effects**. If the selected QTL displays no significant interaction with environment, then a test is performed of whether the dominance effect has a significant contribution in the combined QTL effect. If the selected QTL has a significant QTL×E interaction, a test is performed of whether the dominance-by-environment interaction has a significant contribution in the combined QTL×E interaction.

In this example, as only additive effects being modelled, Step 3 is not relevant.

The first section of the output reports the QTLs retained in the model and whether they show a significant interaction with environment.

```
QTL backward selection for loci in multiple environment trials
================================================================


Summary
=======

Trait: yld_new
--------------

       Term  env_new.pred[112] is removed from the model


Significant terms
-----------------
```

```
 Locus       IdLocus Chromosome     Position Interaction
    19          L085          1       141.00           1
    42         C2P36          2        35.90           1
    73          L035          3        55.70           1
   112          L071          4       136.60           0
   159          L043          6       125.00           1
   237        C10P60         10        60.15           1
```

In this case, all six candidates are retained in the model and therefore listed under the heading `Significant terms`. The list presents the usual locus description (number, locus name, chromosome, position), plus an extra line that contains an indicator variable under the heading `Interaction`. This logical variable contains a 1 when the QTL×E is significant and a 0 when the QTL×E is not significant. For this example, 5 of the 6 QTLs have a significant QTL×E interaction. However, there is no evidence of an interaction with environment for the QTL on chromosome 4, `L071`, at loci 112. Consequently only the main effect of QTL `L071` will be retained in the final model: the term associated with its interaction with environment, `env_new.pred[112]`, is removed. The notation `pred[L]` refers to the QTL at locus *L*.

Standard REML output is then given for the selected multi-QTL model, including variance component estimates and significance tests of the QTLs.

```
REML variance components analysis
=================================

Response variate:  yld_new
Fixed model:       Constant + env_new + pred[19] + pred[42] + pred[73] +
pred[112] + pred[159] + pred[237] + env_new.pred[19] + env_new.pred[42] +
env_new.pred[73] + env_new.pred[159] + env_new.pred[237]
Random model:      genotype_new.env_new
Number of units:   1688

genotype_new.env_new used as residual term with covariance structure as below

Sparse algorithm with AI optimisation


Covariance structures defined for random model
----------------------------------------------


Covariance structures defined within terms:

Term                    Factor      Model                 Order  No. rows
genotype_new.env_new    genotype_newIdentity                  0       211
                        env_new     FA (covariance)          16         8
```

```
Residual variance model
-----------------------


Term            Factor        Model(order)  Parameter       Estimate      s.e.
genotype_new.env_new                        Sigma2             1.000      fixed
                genotype_new  Identity      -                      -         -
                env_new       FA(1) (covariance form)
                                            g_11               67.16      9.13
                                            g_21               88.04      8.93
                                            g_31               103.8       9.1
                                            g_41               28.51      4.77
                                            g_51               16.99      4.39
                                            g_61               106.7      13.0
                                            g_71               68.90      8.36
                                            g_81               97.76      9.67
                                            psi_1             11665.      1258.
                                            psi_2              9445.      1136.
                                            psi_3              8483.      1159.
                                            psi_4              3409.       356.
                                            psi_5              3042.       307.
                                            psi_6             22717.      2519.
                                            psi_7              9407.      1043.
                                            psi_8             10956.      1336.



Tests for fixed effects
-----------------------


Sequentially adding terms to fixed model

Fixed term             Wald statistic  n.d.f.   F statistic  d.d.f.    F pr
env_new                       7517.95       7      1061.16    415.7  <0.001
pred[19]                         0.08       1         0.08    293.1   0.773
pred[42]                         4.28       1         4.28    293.1   0.039
pred[73]                        11.20       1        11.20    293.1  <0.001
pred[112]                       14.14       1        14.14    293.1  <0.001
pred[159]                        0.11       1         0.11    293.1   0.736
pred[237]                        5.03       1         5.03    293.1   0.026
env_new.pred[19]                83.18       7        11.74    415.7  <0.001
env_new.pred[42]                26.77       7         3.78    415.7  <0.001
env_new.pred[73]                30.36       7         4.28    415.7  <0.001
env_new.pred[159]               31.44       7         4.44    415.7  <0.001
env_new.pred[237]               49.75       7         7.02    415.7  <0.001


Dropping individual terms from full fixed model

Fixed term             Wald statistic  n.d.f.   F statistic  d.d.f.    F pr
pred[112]                       13.09       1        13.09    293.1  <0.001
```

```
env_new.pred[19]                          82.22     7        11.61  415.7  <0.001
env_new.pred[42]                          29.24     7         4.13  415.7  <0.001
env_new.pred[73]                          25.42     7         3.59  415.7  <0.001
env_new.pred[159]                         30.01     7         4.24  415.7  <0.001
env_new.pred[237]                         49.75     7         7.02  415.7  <0.001
```

```
* MESSAGE: denominator degrees of freedom for approximate F-tests are calculated
using    algebraic   derivatives   ignoring   fixed/boundary/singular   variance
parameters.
```

Finally, a summary is printed including the estimated QTL effects in the different environments, their standard errors, test of significance, and the estimated locations of the QTLs with 95% confidence intervals. The column `High value allele` indicates which parent provides the high yielding allele (superior allele) for each QTL in each environment. For example, the QTL on chromosome 1 (`L085`) has a larger effect in environment `SS92a` (72.0, parent 2) than environment `IS92a` (55.3, parent 2). In environment `SS92a`, replacing the parent 1 allele by the parent 2 allele at this QTL is expected to increase yield by 72.0 kg/plot (s.e. = 12.19 kg/plot), whereas in environment `IS92a` the expected increase is only 55.3 kg/plot (s.e. = 13.44 kg/plot). Note, that the effect in `HN96b` of 37.5 kg/plot (s.e. = 13.04 kg/plot) is associated with the allele from parent 1 as the high value one (an example of a cross-over interaction).

The lower and upper 95% confidence intervals limits for the estimated locations of the QTLs are given by `CI_LL` and `CI_UL`, respectively.

When backward selection is used to determine the final model, if the QTL×E interaction is non-significant, the QTL effect is constrained to be constant across all environments studied. In this example, QTL×E interactions are significant at all but `L071`. For QTL `L071`, replacing a parent 2 allele by a parent 1 allele is expected to increase yield by 16.3 kg/plot (s.e. = 4.53 kg/plot) in all 8 environments.

```
Estimation of QTL effects from a multi-environment trial
========================================================



REML variance components analysis
=================================

Response variate:  yld_new
Fixed model:       Constant + env_new + env_new.pred[19] + env_new.pred[42] +
env_new.pred[73] + pred[112] + env_new.pred[159] + env_new.pred[237]
Random model:      genotype_new.env_new
Number of units:   1688
```

genotype_new.env_new used as residual term with covariance structure as below

Sparse algorithm with AI optimisation

Covariance structures defined for random model

Covariance structures defined within terms:

Term	Factor	Model	Order	No. rows
genotype_new.env_new	genotype_newIdentity		0	211
	env_new	FA (covariance)	16	8

Residual variance model

Term	Factor	Model(order)	Parameter	Estimate	s.e.
genotype_new.env_new			Sigma2	1.000	fixed
	genotype_new	Identity	–	–	–
	env_new	FA(1) (covariance form)			
			g_11	67.16	9.13
			g_21	88.04	8.93
			g_31	103.8	9.1
			g_41	28.51	4.77
			g_51	16.99	4.39
			g_61	106.7	13.0
			g_71	68.90	8.36
			g_81	97.76	9.67
			psi_1	11665.	1258.
			psi_2	9445.	1136.
			psi_3	8483.	1159.
			psi_4	3409.	356.
			psi_5	3042.	307.
			psi_6	22717.	2519.
			psi_7	9407.	1043.
			psi_8	10956.	1336.

Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
env_new	7517.94	7	1061.16	415.7	<0.001
env_new.pred[19]	83.26	8	10.25	400.8	<0.001
env_new.pred[42]	31.05	8	3.82	400.8	<0.001

```
env_new.pred[73]                        41.56      8       5.12   400.8  <0.001
pred[112]                               14.14      1      14.14   293.1  <0.001
env_new.pred[159]                       31.56      8       3.89   400.8  <0.001
env_new.pred[237]                       54.78      8       6.75   401.0  <0.001


Dropping individual terms from full fixed model

Fixed term                 Wald statistic  n.d.f.   F statistic  d.d.f.    F pr
env_new.pred[19]                        82.23      8      10.13   400.8  <0.001
env_new.pred[42]                        35.51      8       4.37   400.8  <0.001
env_new.pred[73]                        35.48      8       4.37   400.8  <0.001
pred[112]                               13.09      1      13.09   293.1  <0.001
env_new.pred[159]                       30.21      8       3.72   400.8  <0.001
env_new.pred[237]                       54.78      8       6.75   401.0  <0.001


* MESSAGE: denominator degrees of freedom for approximate F-tests are calculated
using   algebraic   derivatives   ignoring  fixed/boundary/singular   variance
parameters.




Summary
=======


Trait: yld_new
Population type: F2
Number of genotypes: 211
Number of environments: 8
Number of linkage groups: 10
Number of markers: 122
Variance-covariance model: FA

List of QTLs
============


Locus no.   Locus name          Linkage group  Position   -log10(P)      QTLxE
       19   L085                            1    141.00      13.760         yes
       42   C2P36                           2     35.90       4.665         yes
       73   L035                            3     55.70       4.661         yes
      112   L071                            4    136.60       3.527          no
      159   L043                            6    125.00       3.712         yes
      237   C10P60                         10     60.15       8.313         yes


QTL (Locus name): L085
======================


Location: linkage group 1 position 141
```

Environment	Effect	High value allele	s.e.	P	%Expl. var.	CI_LL	CI_UL
HN96b	37.466	Parent1	13.035	0.004	3.1	124.55	157.45
IS92a	55.323	Parent2	13.440	0.000	7.2	124.55	157.45
IS94a	56.192	Parent2	14.226	0.000	7.2	124.55	157.45
LN96a	0.117	Parent2	6.660	0.986	0.0	*	*
LN96b	1.577	Parent2	5.915	0.790	0.0	*	*
NS92a	63.762	Parent2	18.925	0.001	5.2	124.55	157.45
SS92a	72.019	Parent2	12.193	0.000	15.1	124.55	157.45
SS94a	27.543	Parent2	14.678	0.061	1.7	*	*

QTL (Locus name): C2P36
=======================

Location: linkage group 2 position 35.9

Environment	Effect	High value allele	s.e.	P	%Expl. var.	CI_LL	CI_UL
HN96b	19.080	Parent2	15.757	0.226	0.8	*	*
IS92a	48.550	Parent1	16.247	0.003	5.5	0.00	198.50
IS94a	3.674	Parent1	17.196	0.831	0.0	*	*
LN96a	2.377	Parent2	8.051	0.768	0.1	*	*
LN96b	22.827	Parent2	7.151	0.001	6.7	0.00	198.50
NS92a	55.737	Parent1	22.877	0.015	4.0	0.00	198.50
SS92a	18.973	Parent1	14.739	0.198	1.0	*	*
SS94a	13.281	Parent1	17.744	0.454	0.4	*	*

QTL (Locus name): L035
======================

Location: linkage group 3 position 55.7

Environment	Effect	High value allele	s.e.	P	%Expl. var.	CI_LL	CI_UL
HN96b	57.120	Parent2	13.378	0.000	7.2	0.00	225.00
IS92a	15.282	Parent2	13.794	0.268	0.5	*	*
IS94a	25.723	Parent2	14.601	0.078	1.5	*	*
LN96a	0.503	Parent1	6.836	0.941	0.0	*	*
LN96b	22.604	Parent2	6.072	0.000	6.5	0.00	225.00
NS92a	6.785	Parent1	19.424	0.727	0.1	*	*

166

```
SS92a              19.177    Parent2   12.515    0.125     1.1        *          *
SS94a               6.379    Parent2   15.065    0.672     0.1        *          *


QTL (Locus name): L071
======================


Location: linkage group 4 position 136.6
----------------------------------------
```

Environment	Effect	High value allele	s.e.	P	%Expl. var.	CI_LL	CI_UL
HN96b	16.369	Parent1	4.525	0.000	0.6	0.00	167.50
IS92a	16.369	Parent1	4.525	0.000	0.6	0.00	167.50
IS94a	16.369	Parent1	4.525	0.000	0.6	0.00	167.50
LN96a	16.369	Parent1	4.525	0.000	3.1	0.00	167.50
LN96b	16.369	Parent1	4.525	0.000	3.4	0.00	167.50
NS92a	16.369	Parent1	4.525	0.000	0.3	0.00	167.50
SS92a	16.369	Parent1	4.525	0.000	0.8	0.00	167.50
SS94a	16.369	Parent1	4.525	0.000	0.6	0.00	167.50

```
QTL (Locus name): L043
======================


Location: linkage group 6 position 125
--------------------------------------
```

Environment	Effect	High value allele	s.e.	P	%Expl. var.	CI_LL	CI_UL
HN96b	25.264	Parent1	12.629	0.045	1.4	6.20	158.70
IS92a	4.852	Parent2	13.022	0.709	0.1	*	*
IS94a	9.245	Parent2	13.783	0.502	0.2	*	*
LN96a	15.021	Parent1	6.452	0.020	2.6	6.20	158.70
LN96b	16.484	Parent2	5.731	0.004	3.5	6.20	158.70
NS92a	43.761	Parent1	18.337	0.017	2.5	6.20	158.70
SS92a	0.061	Parent2	11.814	0.996	0.0	*	*
SS94a	0.341	Parent1	14.222	0.981	0.0	*	*

```
QTL (Locus name): C10P60
========================


Location: linkage group 10 position 60.15
-----------------------------------------
```

Environment	Effect	High value allele	s.e.	P	%Expl. var.	CI_LL	CI_UL
HN96b	92.604	Parent2	13.622	0.000	19.0	48.08	72.22
IS92a	56.526	Parent2	14.045	0.000	7.5	48.08	72.22
IS94a	43.053	Parent2	14.866	0.004	4.2	48.08	72.22
LN96a	9.714	Parent2	6.964	0.163	1.1	*	*
LN96b	7.852	Parent2	6.187	0.204	0.8	*	*
NS92a	47.614	Parent2	19.776	0.016	2.9	48.08	72.22
SS92a	13.875	Parent2	12.743	0.276	0.6	*	*
SS94a	61.011	Parent2	15.339	0.000	8.2	48.08	72.22

We've also requested several plots to be produced (see Figure 7.5). The first, Figure 7.6, plots the significant QTLs on a genetic map. QTLs identified as having a significant QTL×E interaction are plotted in blue, and those with a non-significant interaction in red.
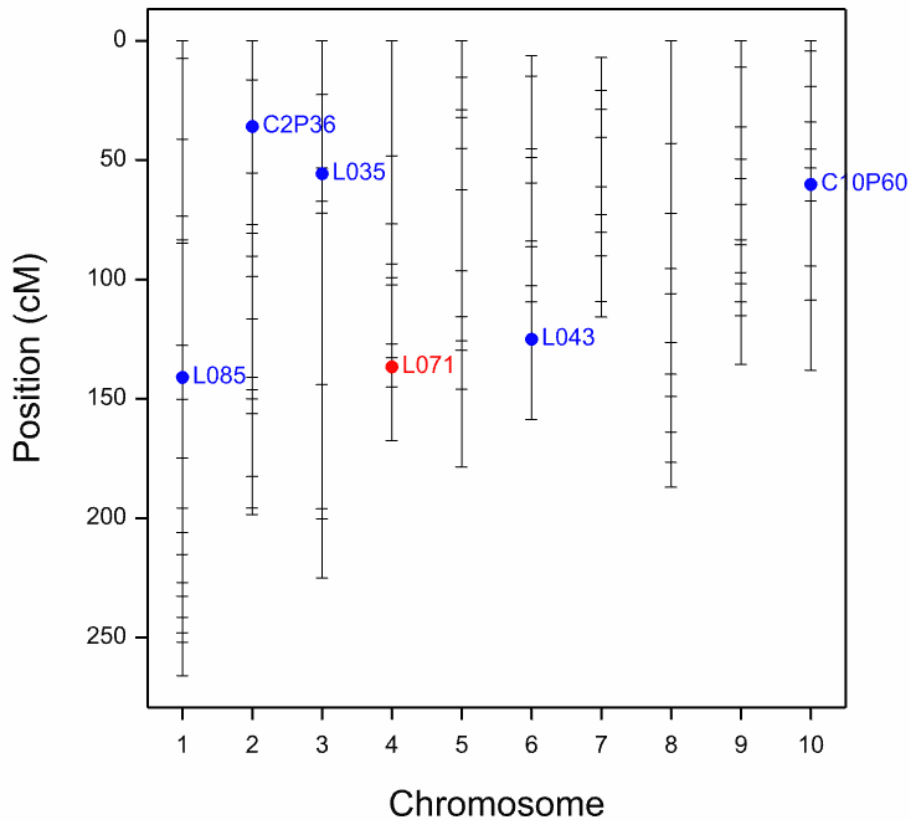
Figure 7.6: Genetic map for the CIMMYT maize trials showing detected QTLs for yield.

   The option Bar charts of additive effects produces two bar charts. The first plots the QTL additive effect for each QTL within each environment (Figure 7.7a). The second plots the QTL additive effect for each environment within the QTLs (Figure 7.7b). These graphs are useful for visualising the size of the QTL effect in each environment and the nature of the QTL×E interaction. For example, cross-over interactions (see Section 4.1.1) can be identified in Figure 7.7b for loci with large bars on either side of the origin (e.g. `L085`). Loci exhibiting divergent (or convergent) interactions are identified by bars pointing in the same direction, but with large differences in magnitude (e.g. `C10P60`). From Figure 7.7a, environments with bars near the origin elicit only weak interactive forces (e.g. `LN96a`), whereas those with bars far from the origin elicit strong interactive forces (e.g. `NS92a`).

a)

b)
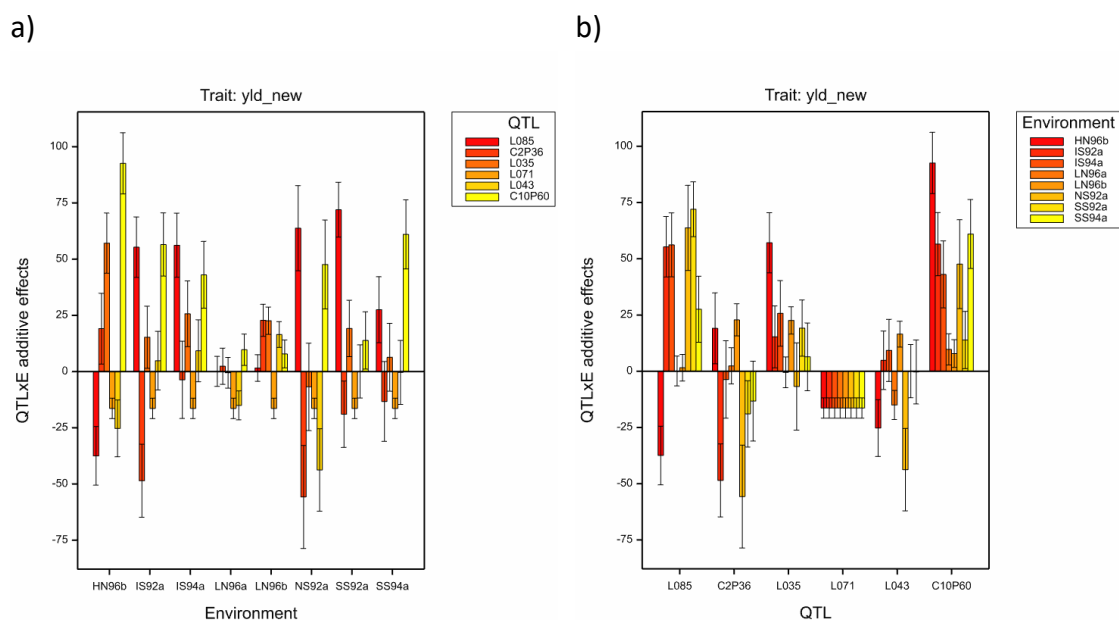
 Figure 7.7: QTL additive effects for each a) QTL within each environment and b) environment within each QTL.

7.3   Multiple trait single environment

Conceptually there are no major differences between multiple trait (single environment) and multiple environment (single trait) QTL analyses. The latter can be viewed as a type of multiple-trait analysis, where the same trait observed in different environments is regarded as different traits. The key motivation for including a separate section on multi-

trait QTL detection is to emphasize that, by default, the variance-covariance structure is always modelled using the unstructured form. This is because the multiple traits under consideration may be rather different in terms of scale and variation, making many of the variance-covariance models difficult to interpret or even meaningless. The only variance-covariance model that is always meaningful in multi-trait QTL analysis is the unstructured model.

The statistical models for multi-trait QTL detection are analogous to those presented in Section 7.2 for multiple environments. However, the environment effects (E_j, $j = 1$, …, m) are replaced by trait effects (T_k, $k = 1$, …, K). For example, the final multi-QTL model (with additive effects only) is expressed as:

$$y_{ik} = \mu + T_k + \sum_{q \in Q} \alpha_{kq} x_{iq} + \varepsilon_{ik} + e_{ik} \qquad \text{Equation 4}$$

where
y_{ik} is the trait k mean for genotype i
μ is the overall mean
T_k is the trait k main effect
Q is the set of QTLs, $q = 1$, …, Q
α_{kq} is the effect of QTL q for trait k.
x_{iq} is the genetic predictor of QTL q for genotype i
ε_{ik} is the genetic residual of trait k for genotype i (or residual if unit errors are omitted)
e_{ik} is the unit error of trait k for genotype i.

The residuals, ε_{ik}, representing the unexplained genotype and trait effects, are assumed be Normally distributed with mean 0 and variance-covariance structure VCOV(ε_{ik}). VCOV is modelled explicitly using unstructured model, with `Genotype` fitted as a random term. The QTL (α_{kq}, $k = 1$, …, K, $q = 1$, …, Q) and trait (T_k, $k = 1$, …, K) effects are fitted as fixed effects.

The unit errors, e_{ik}, represent the uncertainty on the trait means (see Section 3.2). However, if estimates of the unit errors are no available, e_{ik} and and ε_{ik}, cannot be separately estimated.

We illustrate QTL analysis for multi-trait, single environment data sets using the Steptoe-Morex barley trial data (described in Section 1.3.1). The marker and map information for this double haploid population are held in Flapjack files `SxM_geno.txt` and `SxM_map.txt`, respectively. Import and inspect the marker and map data (referring to Sections 2.1.2.1 and 2.4.2.3). Remember to specify the population type as double-haploid (`DH1`). Two phenotypic traits were measured: yield (`yield`) and heading date

(`heading`). The trait means are held in file `SxM_pheno.csv`. Use the Summary Statistics Between Traits menu to explore the correlations between the trait means (Section 2.4.1.3).

The Multi-trait Linkage Analysis window (Figure 7.8) can be accessed by either:

- Stats | QTLs (Linkage/Association) | QTL Analysis | Multi-trait Linkage Analysis (Single Environment); or,
- in the QTL Data View via the shortcut QTL analysis | Multi-trait Linkage Analysis (Single Environment).

Genstat will automatically populate the input fields using data from the QTL Data Space. The menu is similar to that for a single trait - single environment analysis except now multiple traits are entered in the Quantitative trait means: box (Figure 7.8). If unit errors are available they should be included in the analysis using the Include unit errors: field accessed via the Options button (refer to Figure 6.7). For multi-trait QTL analysis, any unit errors stored within the QTL Data Space for the different traits will be used. Refer to Chapter 6 for further details on the input fields and Options.
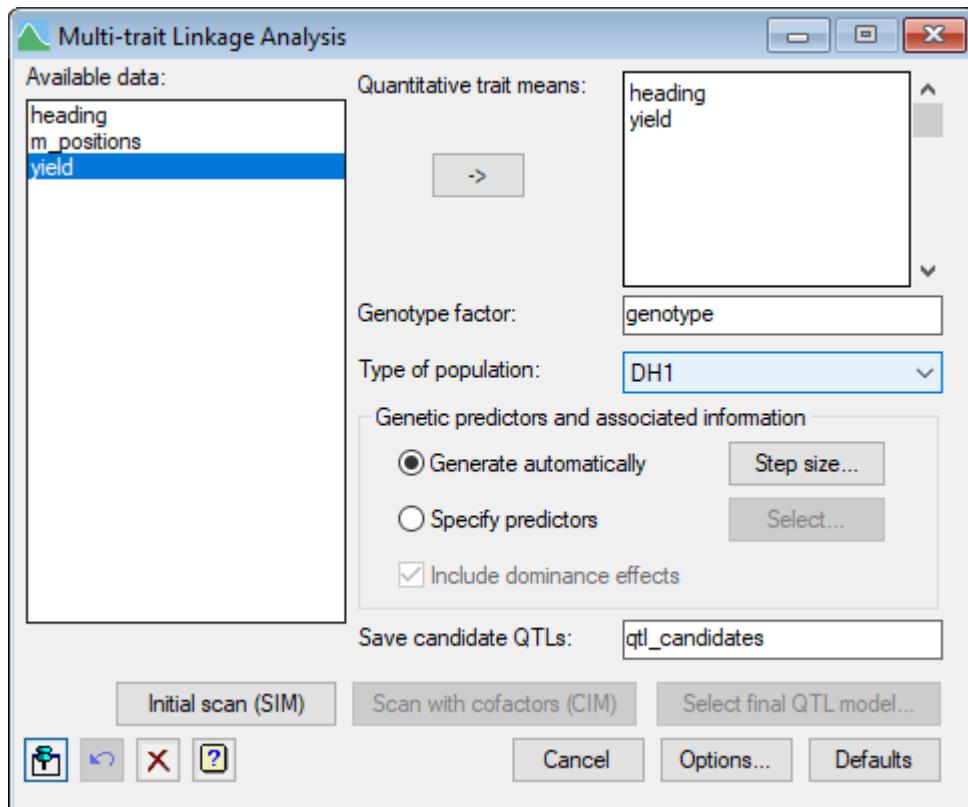
Figure 7.8: Menu for performing a multi-trait, single environment QTL analysis on the quantitative traits `yield` and `heading` from the Steptoe-Morex barley trial.

The process for detecting QTLs is the same as for the single trait case (demonstrated in Chapter 6 and Section 7.2):

1) candidate QTLs are identified using SIM (or marker regression), followed by one or more rounds of CIM;
2) a final multi-QTL model is selected from the set of candidate QTLs.

However, in contrast to the multiple environment case, partitioning QTL effects into main effects and QTL by trait (QTL×T) interactions is rarely meaningful, as the traits under consideration are usually measured in different units.

In this example, after an initial genome-wide scan by SIM and five rounds of CIM (using step size of 10cM), 5 candidate QTLs are identified:

```
The following candidates have been selected

    Locus   IdLocus  Chromosome    Position
       31      abg2           2        41.2
       37    abc162           2        73.5
       61   abg703a           3        83.6
       94    C4P104           4       103.8
      158       ale           7        68.2
```

The profile plot, Figure 7.9, consists of two frames. The upper frame profiles the p-values (on the -$\log_{10}$ scale) against linkage group position. Values above the threshold (red line) are indications that at least one of the traits is affected by a QTL at that chromosome position. In the lower frame you can read which of the traits are affected by a particular candidate QTL. The traits are listed on the Y-axis, with coloured points aligning to positions where the significance test for the QTL effect (here, additive) is larger than threshold. The points are coloured according to which parent provided the allele for a high value (i.e. the superior allele); blue denotes that the high value allele originates from parent 1 and red from parent 2. The intensity of the colour represents the magnitude of the effect.

In this example, two candidate QTLs are identified on chromosome 2 (`abg2` and `abc162`) - located at the two peaks exceeding threshold. The first, `abg2`, affects both traits, `yield` and `heading`, with the allele provided by parent 2 associated with a high value for both. In addition, the dark red colouring indicates that `abg2` has a large effect on both traits. The second, `abc162`, located at the lower peak, only affects `heading`, with high `heading` values associated with the allele coming from parent 1.
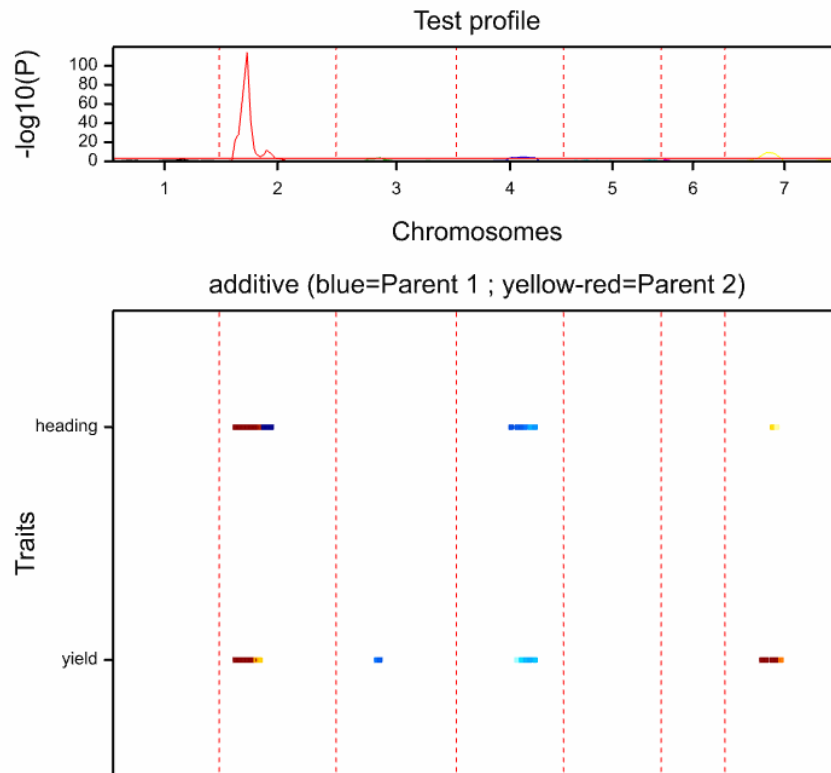
Figure 7.9: Profile plot of the set of candidate QTLs following 1 round of SIM and 5 rounds of CIM, for Steptoe-Morex `yield` and `heading` data. Step size = 10cM.

After a set of candidate QTLs has been defined, the final multi-QTL model can be selected. Click on the Select final QTL model button to open the Select Final Model (Multi-trait) window. The options are the same as described for multiple environment analysis, except the variance-covariance model is set to unstructured (see Section 7.2). The list of candidate QTLs is automatically taken from the last genome-wide search, but you can modify this if you wish using the Candidate QTLs button.

By default backward selection (with significance level 0.05) is used to select the final QTL model. In contrast to the multiple environment case, non-significant QTL by trait (QTL×T) interactions are always retained in the final model: as the traits are usually measured in different units, and likely differ in terms of scale and variation, simplification to QTL main effects is rarely meaningful. (Refer to Section 7.2 for a description of backward selection in a multiple environment analysis).

Click Run to select and fit the final multi-QTL model.

The first section of the output summarizes the results from backward selection. In this example, all five candidate QTLs are significant and therefore included in the final multi-QTL model. Note: the output also reports a non-significant QTL×T interaction at locus 94 (`C4P104`), term `_traitgroup.pred[94]`. However, as our two traits are measured in different units, and weren't standardized prior to analysis, tests of the QTL×T interactions are not of interest.

```
QTL backward selection for loci in multi-trait trials
=====================================================


Summary
=======

Trait: _traits
--------------

      Term  _traitgroup.pred[94] is removed from the model


Significant terms
-----------------

  Locus      IdLocus Chromosome    Position Interaction
     31         abg2          2       41.20           1
     37       abc162          2       73.50           1
     61       abg703a          3       83.60           1
     94        C4P104          4      103.77           0
    158          ale          7       68.20           1
```

The second section of output summarizes the fit of the final multi-QTL model, including the estimated QTL effects for each trait. Note: because QTL main effects across multiple traits are not meaningful, the QTL×T interaction for `C4P104` (locus 94) is retained in the final model.

The significance of the QTL effect for each trait can be assessed using `P`, the p-value obtained by comparing $(\texttt{Effect})^2/(\texttt{s.e.})^2$ to a chi-square distribution on 1 df. The column `High value allele` indicates which parent provides the allele corresponding to the higher phenotypic response (i.e. superior allele). For example, the QTL detected on chromosome 2, `abg2`, has a significant additive effect on both `heading` ($P = 0.000$) and `yield` ($P = 0.000$). For both traits, the high value allele is provided by the Steptoe parent. Replacing a Morex allele with a Steptoe allele is expected to increase `heading` by 0.950 days (s.e. = 0.042 days) and `yield` by 0.446 ton/ha (s.e. = 0.068 ton/ha). In comparison, the QTL detected on chromosome 3, `abg703a`, only significantly affects

yield. Here, replacing a Morex allele with a Steptoe allele is expected to decrease yield by 0.263 ton/ha (s.e. = 0.065 ton/ha).

```
Estimation of QTL effects from a multi-trait trial
==================================================


REML variance components analysis
=================================

Response variate:  newy
Fixed  model:            Constant  +  _traitgroup  +  _traitgroup.pred[31]  +
_traitgroup.pred[37]   +   _traitgroup.pred[61]   +   _traitgroup.pred[94]   +
_traitgroup.pred[158]
Random model:       _genotypes._traitgroup
Number of units:   300

_genotypes._traitgroup used as residual term with covariance structure as below

Sparse algorithm with AI optimisation


Covariance structures defined for random model
----------------------------------------------

Covariance structures defined within terms:

Term                     Factor     Model                         Order  No. rows
_genotypes._traitgroup   _genotypes Identity                          0       150
                         _traitgroup Unstructured                     1         2


Residual variance model
-----------------------

Term            Factor        Model(order)  Parameter       Estimate      s.e.
_genotypes._traitgroup                      Sigma2             1.000     fixed
                _genotypes    Identity      -                      -         -
                 traitgroup   Unstructured  v_11              0.2223    0.0262
                                            v_21             0.05454   0.03051
                                            v_22              0.5897    0.0695


Tests for fixed effects
-----------------------

Sequentially adding terms to fixed model

Fixed term              Wald statistic  n.d.f.  F statistic  d.d.f.   F pr
_traitgroup                       0.00       1         0.00   144.0  1.000
```

```
_traitgroup.pred[31]              458.14      2          227.48   143.0  <0.001
_traitgroup.pred[37]               57.13      2           28.37   143.0  <0.001
_traitgroup.pred[61]               18.73      2            9.30   143.0  <0.001
_traitgroup.pred[94]               21.07      2           10.46   143.0  <0.001
_traitgroup.pred[158]              41.77      2           20.74   143.0  <0.001


Dropping individual terms from full fixed model

Fixed term              Wald statistic  n.d.f.   F statistic  d.d.f.    F pr
_traitgroup.pred[31]            524.78      2          260.57   143.0  <0.001
_traitgroup.pred[37]             52.90      2           26.27   143.0  <0.001
_traitgroup.pred[61]             17.30      2            8.59   143.0  <0.001
_traitgroup.pred[94]             20.34      2           10.10   143.0  <0.001
_traitgroup.pred[158]            41.77      2           20.74   143.0  <0.001


* MESSAGE: denominator degrees of freedom for approximate F-tests are calculated
using   algebraic   derivatives   ignoring   fixed/boundary/singular   variance
parameters.



Summary
=======


Population type: DH1
Number of genotypes: 150
Number of traits: 2
Number of linkage groups: 7
Number of markers: 116
Variance-covariance model: UNSTRUCTURED


List of QTLs
============



Locus no.   Locus name         Linkage group   Position    -log10(P)     QTLxT
       31   abg2                          2       41.20      113.955       yes
       37   abc162                        2       73.50       11.487       yes
       61   abg703a                       3       83.60        3.757       yes
       94   C4P104                        4      103.77        4.418       yes
      158   ale                           7       68.20        9.070       yes


QTL (Locus name): abg2
======================



Location: linkage group 2 position 41.2
---------------------------------------



Trait            Effect High value     s.e.        P    %Expl.   CI_LL    CI_UL
```

		allele			var.		
heading	0.950	Steptoe	0.042	0.000	90.3	26.09	56.31
yield	0.446	Steptoe	0.068	0.000	19.9	26.09	56.31

QTL (Locus name): abc162
========================

Location: linkage group 2 position 73.5

Trait	Effect	High value allele	s.e.	P	%Expl. var.	CI_LL	CI_UL
heading	0.306	Morex	0.042	0.000	9.3	0.00	183.10
yield	0.102	Morex	0.069	0.137	1.0	*	*

QTL (Locus name): abg703a
=========================

Location: linkage group 3 position 83.6

Trait	Effect	High value allele	s.e.	P	%Expl. var.	CI_LL	CI_UL
heading	0.064	Morex	0.040	0.112	0.4	*	*
yield	0.263	Morex	0.065	0.000	6.9	16.30	205.00

QTL (Locus name): C4P104
========================

Location: linkage group 4 position 103.8

Trait	Effect	High value allele	s.e.	P	%Expl. var.	CI_LL	CI_UL
heading	0.180	Morex	0.044	0.000	3.3	1.40	168.40
yield	0.184	Morex	0.072	0.011	3.4	1.40	168.40

QTL (Locus name): ale
=====================

```
Location: linkage group 7 position 68.2
---------------------------------------
```

Trait	Effect	High value allele	s.e.	P	%Expl. var.	CI_LL	CI_UL
heading	0.061	Steptoe	0.039	0.119	0.4	*	*
yield	0.409	Steptoe	0.064	0.000	16.8	4.70	191.20

The significant QTLs are plotted on a genetic map in Figure 7.10.
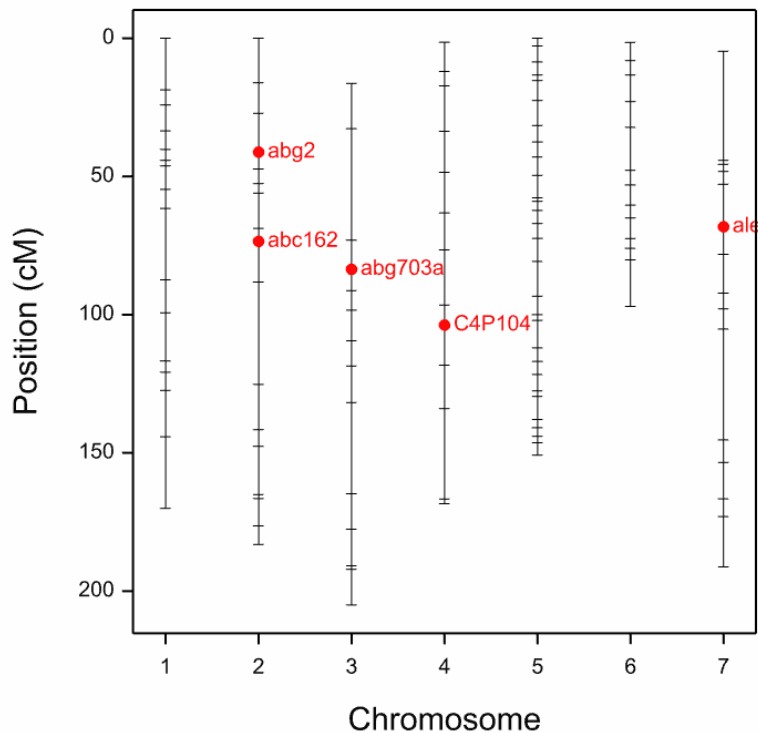
Figure 7.10: Genetic map for the Steptoe-Morex data showing detected QTLs for `yield` and `heading`.

Two bar charts are given in Figure 7.11 - these are useful for visualising the size and nature of the QTL additive effects. The first plots the QTL additive effect for each QTL within each trait. The second plots the QTL additive effect for each trait within each QTL. In this example, for all QTLs the direction of the additive effect is the same for both traits (Figure 7.11b). However, within a trait, the magnitude differs considerably (Figure 7.11a); for both `yield` and `heading` the largest (and positive) effect is associated with

`abg2`. QTL `ale` also has a large (and positive) effect on `yield`, although not on `heading`.
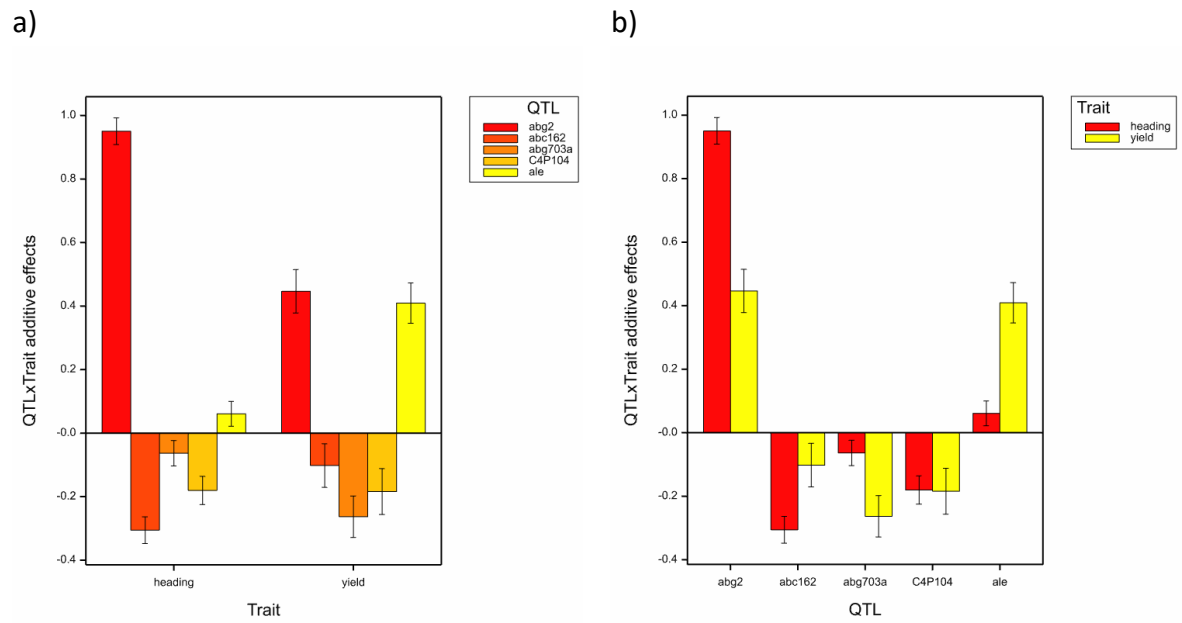
Figure 7.11: QTL additive effects for each a) QTL within each trait and b) trait within each QTL.

7.4 References

Li, J., & Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, **95,** 221-227.

8 Linkage analysis: cross pollinated populations

8.1 QTL linkage analysis in Genstat

In Genstat, linkage analysis is also possible for a cross-pollinated (CP) population derived from the cross of two heterozygous parents. This chapter provides a very brief overview of Genstat's QTL facilities for cross-pollinated populations. A more comprehensive description will be made available in a later edition of the manual.

For cross-pollinated populations, additive, 2nd parent additive and dominance genetic predictors can be calculated. These are obtained using the Stats | QTLs (Linkage/Association) | Genotypic Analysis | Calculate Genetic Predictors window (Figure 8.1). Note, the Calculate Genetic Predictors window can also be accessed from the QTL analysis menus by checking Generate automatically (see Figure 6.5, Figure 7.2, or Figure 7.8). By default the additive, 2nd parent additive and dominance genetic predictors are saved in structures `gp_additive`, `gp_additive2` and `gp_dominance`, respectively.
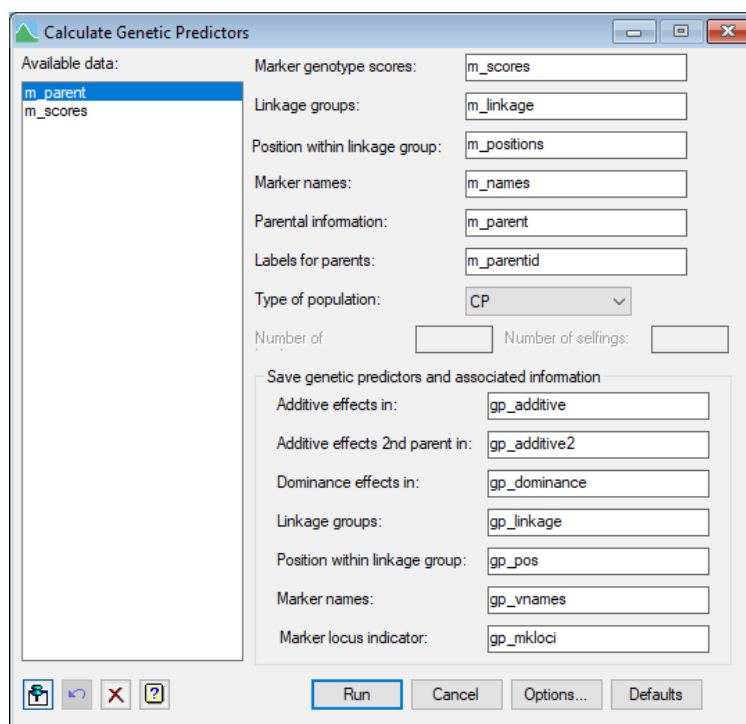
Figure 8.1: Calculating additive, 2nd parent additive and dominance genetic predictors for a cross-pollinated population using the Calculate Genetic Predictors menu.

QTL linkage analysis for a cross-pollinated population then proceeds as for an inbred population (refer to Chapters 6 and 7). Namely: first a putative QTL detection step, where the genome is searched for candidate QTLs (using marker regression or SIM, and preferably, CIM), followed by selection of a final multi-QTL model from the set of candidate QTLs.

The models for detecting QTLs, described previously in Chapters 6 and 7, are readily modified to accommodate dominance effects and second parent additive effects. For example, the marker regression model of Section 6.1.2.1 (single trait, single environment) extended to include both dominance and second parent additive effects is expressed as:

$$y_i = \mu + \alpha^a x_i^a + \alpha^{a2} x_i^{a2} + \alpha^d x_i^d + \varepsilon_i + e_i$$

where

y_i is the trait mean for genotype i

α^a is the additive QTL effect at the position being tested

x_i^a is the additive genetic predictor for genotype i at the position being tested

α^{a2} is the additive QTL effect of the second parent at the position being tested

x_i^{a2} is the additive genetic predictor of the second parent for genotype i at the position being tested

α^d is the dominance QTL effect at the position being tested

x_i^d is the dominance genetic predictor for genotype i at the position being tested

ε_i is the genetic residual for genotype i (or residual if unit errors are omitted)

e_i is the unit error for genotype i.

Note, if estimates of unit error (e_i) are not unavailable, the term is omitted from the model and ε_i now represents the *residual.*

Genstat's menus for a single trait – single environment (Figure 6.5), single trait – multiple environment (Figure 7.2) or multiple trait – single environment (Figure 7.8) QTL linkage analysis allow `CP` to be select in the Type of population: field. When `CP` is selected, the relevant fields to supply data structures for QTL linkage analysis on a cross-pollinated population will be enabled.

9   Association mapping

Association mapping, also known as linkage disequilibrium (LD) mapping, is a method for QTL detection. In contrast to QTL linkage analysis (Chapters 6, 7 and 8), it accommodates broader populations, searching for marker-trait associations in genetically diverse populations.

Widely applied in human genetics, more recently association mapping has gained attention in plant breeding (Zhu *et al*., 2008). As there is no need to develop specific crosses, association mapping can take advantage of existing diverse collections of genotypes. In addition, it can target a broader and more relevant genetic spectrum for plant breeders than conventional QTL mapping.

Association mapping essentially consists of finding marker-trait associations. However, the presence of population structure (i.e. genetic relatedness) can result in "*spurious associations*". That is, where the marker-trait association is not linked to any causative loci. Several strategies have been proposed to account for genetic relatedness, either by stratifying the population into subpopulations (Pritchard *et al*., 2000a,b; Kraakman *et al*., 2004) or including estimates of genetic relatedness between genotypes, i.e. "*kinship*", in the statistical model (Yu *et al*., 2006; Malosetti *et al*., 2007). Alternatively eigenanalysis, a method that approximates the use of the kinship matrix, can be used (Patterson *et al*., 2006). All three approaches are available in Genstat.

This chapter describes association mapping for single trait-single environment data sets with bi-allelic markers. Association mapping for more complex data sets, i.e. from multiple environments and/or with multi-allelic markers, is briefly introduced in Sections 9.5 and 9.6, respectively.

In this chapter you will learn how to:

- investigate population structure by means of eigenanalysis (Section 9.2)
- investigate LD between neighbouring markers on the same chromosome using LD decay plots corrected for genetic relatedness (Section 9.3)
- find marker-trait associations using mixed models that accommodate genetic relatedness using kinship information (Section 9.4.2), eigenanalysis (Section 9.4.3) or *a priori* subpopulation groupings (Section 9.4.4)
- perform association mapping on a single trait-single environment data set with biallelic markers (Section 9.4.5)

Throughout this chapter, association mapping in Genstat is illustrated using the MABDE barley association panel described in Section 1.3.4. Mean yields for each genotype are held in `AMP_Barley_pheno.csv`, along with subpopulation groupings. Marker scores and map information are held in files `AMP_Barley_geno.txt` and `AMP_Barley_map.txt`, respectively. When importing the genotypic data (see Section 2.1) remember to specify the population type as `Association mapping`.

The kinship matrix for the MABDE barley association panel is held in file `AMP_Barley_Kmatrix.txt`. Refer to Section 2.1.3 to load the kinship matrix and the subpopulation information.

9.1   Input data

Three basic data files are necessary to perform association mapping: the map file, the genotype file, and the phenotype file (see Chapter 2).

The genotype file contains genotype by marker (locus) scores. In Genstat, these scores are stored in *m* factors, one for each marker, and held in `m_scores` by default (see Section 2.1.2). Prior to eigenanalysis, LD analysis and association mapping, the scores are converted to variates (genetic predictors) based on allele frequencies. For bi-allelic markers, the most frequent allele is chosen as the "*reference*" allele and the second as the "*variant*" allele. The genetic predictor, $x_{i,m}$, is the number of variant alleles for marker *m*, genotype *i*. For example, the bi-allelic marker with reference allele A and variant allele B, takes the values 0, 1, 2 for AA, AB, and BB, respectively. For multi-allelic markers, refer to Section 9.6.

The phenotypic data contains the quantitative traits (phenotypes) measured on all genotypes in the population. Association mapping in Genstat requires trait means for each genotype. If the raw plot (unit) data are available, read Chapter 3 to obtain trait means.

In association mapping a fourth type of file containing genetic relationship information, either a kinship matrix or subpopulation grouping factor, can also be used (see Section 2.1.3). The kinship matrix contains the coefficient of co-ancestry between all pairs of genotypes in the population. The subpopulation factor groups similar genotypes (using for example, geographic origin, genetic relatedness, etc).

Prior to analysis, it is important to ensure that the ordering of genotypes in the phenotypic, genotypic, and genetic relationship data sets are consistent (refer to Section 2.3.1).

9.2   Investigating population structure

Linkage disequilibrium (LD) is the phenomenon where alleles at different loci are found together in the same gamete more or less frequently that expected based on their frequencies. That is, the "*non-random association of alleles at different loci*" (Flint-Garcia *et al*., 2003). LD is affected by recombination and mutation rates, population size, and selection pressure. LD is the cornerstone of association mapping, where trait-marker relationships are identified from a population of heterogeneously-related individuals.

A crucial aspect of association mapping, and one of the major differences with conventional QTL linkage approaches, is that LD between markers, and between markers and QTLs, can occur even when there is no genetic linkage between them. A major source of LD not related to physical proximity between markers (or QTLs) is genetic relatedness between individuals in the population. Therefore, an important first step in association mapping is to investigate the genetic structure of the population. A popular method is the approach described by Pritchard *et al.* (2000a,b) and implemented in the program STRUCTURE, where subpopulation structure is inferred and individuals are assigned to groups. However, this approach can be computationally intensive. An alternative strategy suggested by Patterson *et al.* (2006), "*eigenanalysis*", uses the scores of the most significant principal components to describe population structure.

Eigenanalysis is a principal components method applied to the matrix of genotype by marker scores (after conversion to variates, see Section 9.1). The method infers the underlying genetic substructure in the population, effectively approximating the kinship matrix. For each marker, missing genetic predictors are replaced by the mean. The Tracy-Widom statistic is used to determine the number of significant principal components.

We investigate the structure of the MABDE barley population, described in Section 1.3.4, using eigenanalysis. The Eigenanalysis window (Figure 9.1) can be accessed either:
- by Stats | QTLs (Linkage/Association) | Genotypic Analysis | Eigenanalysis; or,
- from the QTL Data View via the shortcut Genotypic analysis | Eigenanalysis.
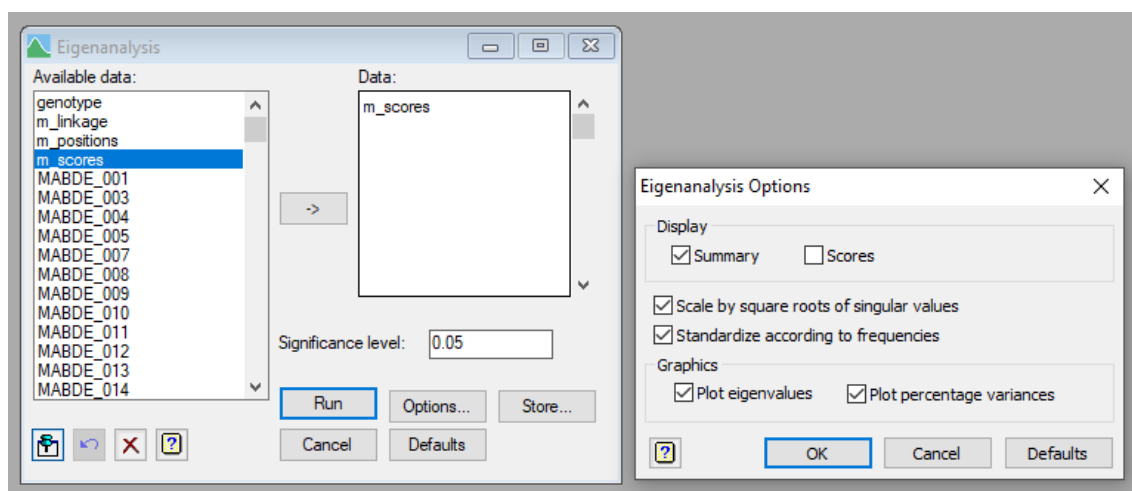
Figure 9.1: Eigenanalysis and Eigenanalysis Options windows.

In the Data: field select the pointer that contains the marker scores (in this example, `m_scores`). The Significance level: for the Tracy-Widom test defaults to 0.05. In the Eigenanalysis Options window (Figure 9.1) you can specify whether to scale the principal component scores by the square root of their singular values (the default) and whether to standardize the scores, prior to analysis, according to frequencies (also the default). The output produced is controlled by the Display and Graphics panes.

The Summary output displays the significant principal components (i.e. `axes`), their associated Tracy-Widom statistic, eigenvalue and percentage variation they explain. The cumulative percentage variance explained is also given. In this example, four significant principal components were found that collectively explain 35.78% of the variation in the MABDE barley marker score data.

```
Detection of significant principal components
=============================================


Summary
=======


Number of significant axes = 4

            Tracy-                            Cumulative
             Widom                %Variance    %variance
   Axis    statistic  Eigenvalue  explained    explained
axis 1       20.92        68843      11.86        11.86
axis 2       23.95        64541      11.12        22.99
axis 3       10.66        42266       7.28        30.27
axis 4        4.13        31977       5.51        35.78
```

The option Scores (not selected here; Figure 9.1) outputs the scores of the significant principal components. Plot eigenvalues graphs the eigenvalues against the number of principal components, useful for understanding how the size of the eigenvalue depreciates (Figure 9.2a). Plot percentage variances graphs the percentage and cumulative percentage of the variance explained against the number of principal components (Figure 9.2b). This plot is useful for visualising how the gains in variation explained diminish as the number of components increase.

The Store button on the Eigenanalysis Options window (Figure 9.1) opens a menu to enable results from an eigenanalysis (Scores, Eigenvalues, Percentage variances and Cumulative percentage variances) to be saved.
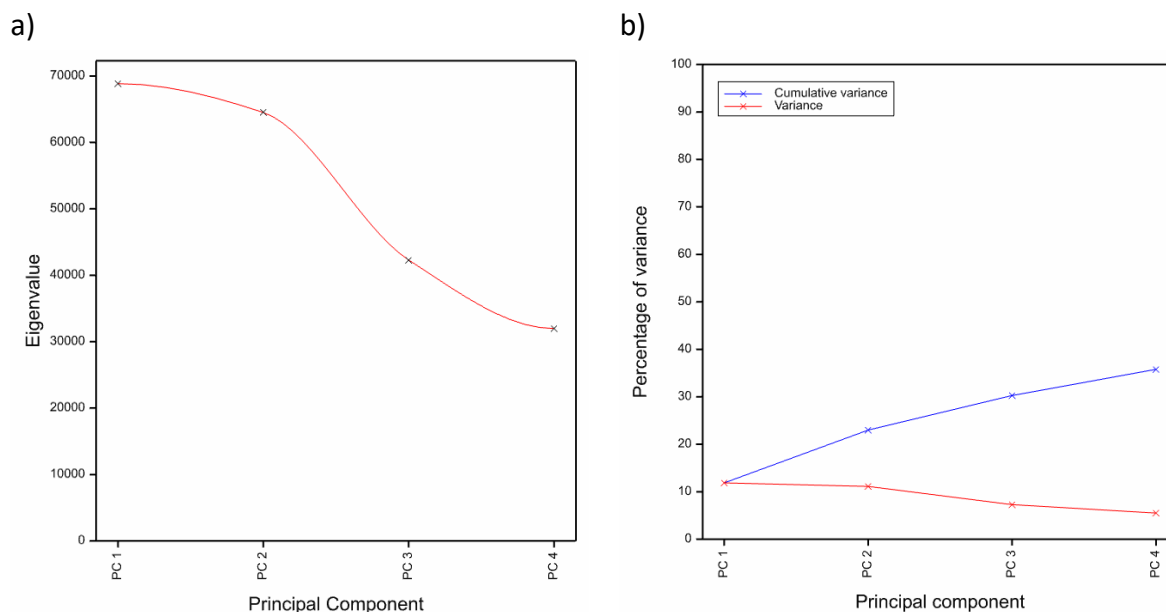
Figure 9.2: Graphical output from an eigenanalysis of the MABDE barley marker score data. Plot of a) eigenvalues and b) variance explained against the number of principal components.

9.3   Investigating LD decay along chromosomes

LD is affected by recombination. An important advantage of association mapping is that it profits from a long history of recombination events in the population. The longer the recombination history, the faster the LD between neighbouring markers will decay. A fast LD decay implies that only tightly linked markers will remain associated to QTLs.

LD decay plots are a useful tool for investigating LD in relation to map distance. As the correlation between alleles at a pair of marker loci is a measure LD, a common approach is to calculate r^2 values (square of the correlation coefficients) between the genetic predictors of markers and plot them against marker distance. However, a major drawback of this approach is that it takes no account of genetic relatedness between individuals in the population. High r^2 values between unlinked markers may be the consequence of genetic relatedness and not of a short physical distance. This leads to an overestimation of LD between markers.

Genstat can be used to calculate LD between markers on the same chromosome whilst accounting for genetic relatedness. The approach fits a linear regression to pairs of

markers, where the genetic predictors of one marker are taken as the response variate and the genetic predictors of the other as the explanatory variate. The information on genetic relatedness is included by either a grouping factor indicating subpopulations of genotypes (for example, based on geographical origin, or groups from STRUCTURE (Pritchard *et al.*, 2000a)), or covariables from the scores of the significant principal components selected by eigenanalysis (see Section 9.2). The association between markers is assessed by the deviance ratio between the models with and without the explanatory genetic predictor variable. From each fitted regression, the r^2 value is stored as the measure of LD between markers on the same chromosome.

The Linkage Disequilibrium (LD) Decay window (Figure 9.3) can be opened either:
- by Stats | QTLs (Linkage/Association) | Genotypic Analysis | Linkage Disequilibrium (LD) Decay; or,
- from the QTL Data View via the shortcut Genotypic analysis | Linkage Disequilibrium (LD) Decay.
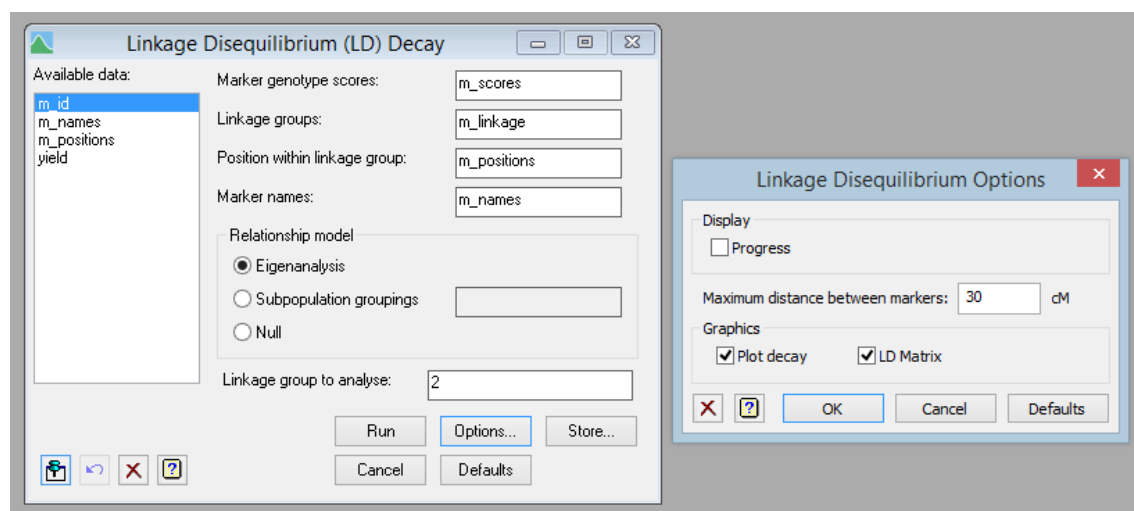
Figure 9.3: Linkage Disequilibrium (LD) Decay and Linkage Disequilibrium Options windows.

Genstat will automatically populate the genotypic input fields using information in the QTL Data Space. In the Relationship model box we specify the approach to account for genetic relatedness. We'll first select Eigenanalysis.

LD is estimated per chromosome (i.e. linkage group), and the chromosome to analyse must be specified in the Linkage group to analyse: field. In this case, we select chromosome 2. Results from an LD analysis (distances and r^2 between markers) can be saved via the menu accessed from the Store button.

Two plots can be requested in the Linkage Disequilibrium Options window (Figure 9.3); Plot decay (the default), a graph of r^2 against marker distance (Figure 9.4a), and LD Matrix, a shade plot of the p-values for the deviance ratios, on the -$\log_{10}$ scale (Figure 9.4b). Although LD is calculated along the whole of the chromosome, LD is expected to decay within relatively short distances. Therefore, in the Linkage Disequilibrium Options (Figure 9.3) you can specify the Maximum distance between markers: for which r^2 is displayed in the decay plot; default 30cM. In this example, LD has decayed substantially by 5cM (Figure 9.4a).

Each cell in the shade plot (Figure 9.4b) represents the LD between a pair of markers on chromosome 2, starting with markers 1 vrs 147 in the lower left corner. The colours represent the strength of the LD between markers: higher $\log_{10}$(p-value), the brighter the red, the stronger the LD. We expect highest LDs near the diagonal, where closely neighbouring markers are plotted.

a)

b)
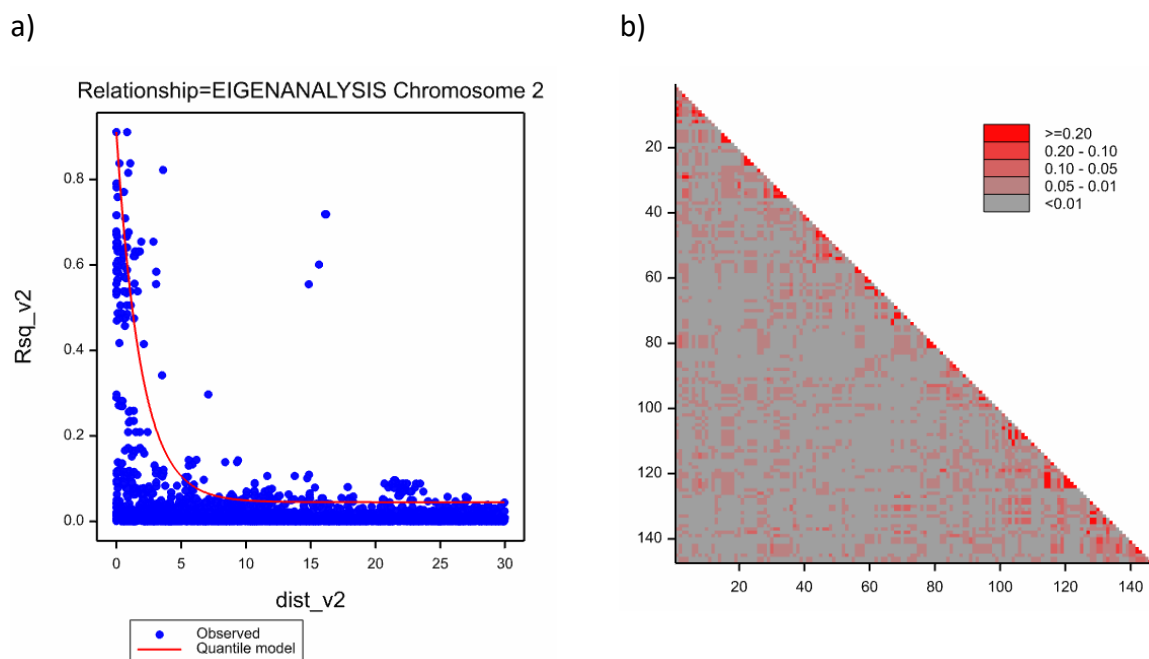
Figure 9.4: Graphical output from an LD analysis of the MABDE barley marker score data for chromosome 2. a) decay plot, b) shade plot. Information on genetic relatedness is included as covariables from eigenanalysis.

Note, LD is not calculated between markers with too many missing values, by default more than 20%. This option cannot be changed via the menus, but can be specified using the `MAX%MISSING` option in the command line.

Genstat's LD analysis can also accommodate an *a priori* population structure by providing a Subpopulation groupings factor (Figure 9.3). For this example, the factor `group` subsets the 179 genotypes of the MABDE barley panel into 5 groups, ranging in size from 14 to 53 genotypes. The decay and shade plots under this alternative specification for the Relationship model are given Figure 9.5a and b, respectively. In this case, the plots are very similar to those produced using the eigenanalysis correction (Figure 9.4), indicating that `group` is capturing similar population structure as the significant principal components.

It is also possible to perform LD analysis without any correction for genetic relatedness by selecting Null as the relationship model. When no correction for genetic relatedness is made, the LD decay plot (Figure 9.5c) reveals high LD not only in physically close markers (i.e. < 5cM) but also between markers physically distant from one another. The shade plot (Figure 9.5d) shows high LD between markers all along the linkage group. This is indicative of population structure. Providing a sensible relationship model clearly diminishes the effect of genetic relatedness on LD, with fewer occurrences of high LD at distances greater than 5cM (Figure 9.4 and Figure 9.5a, b). Once the relationships within the population are accommodated, the majority of high LD values occur where the physical distances between markers are small (i.e. the red clusters near the diagonals in Figure 9.4b and Figure 9.5b).

a)

b)
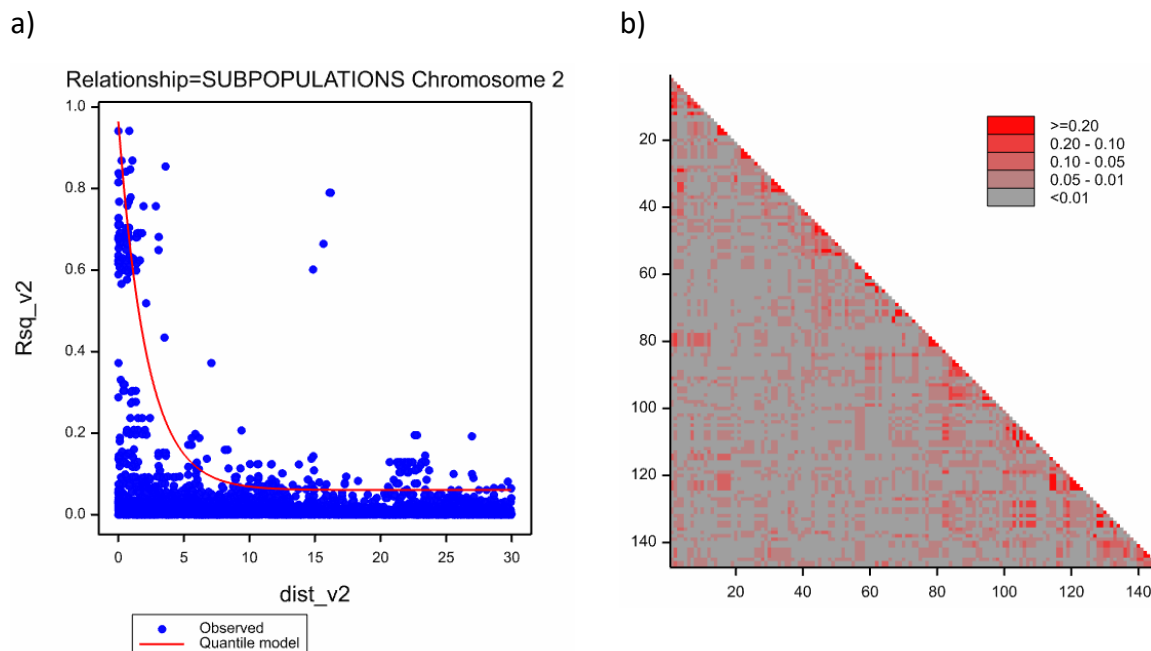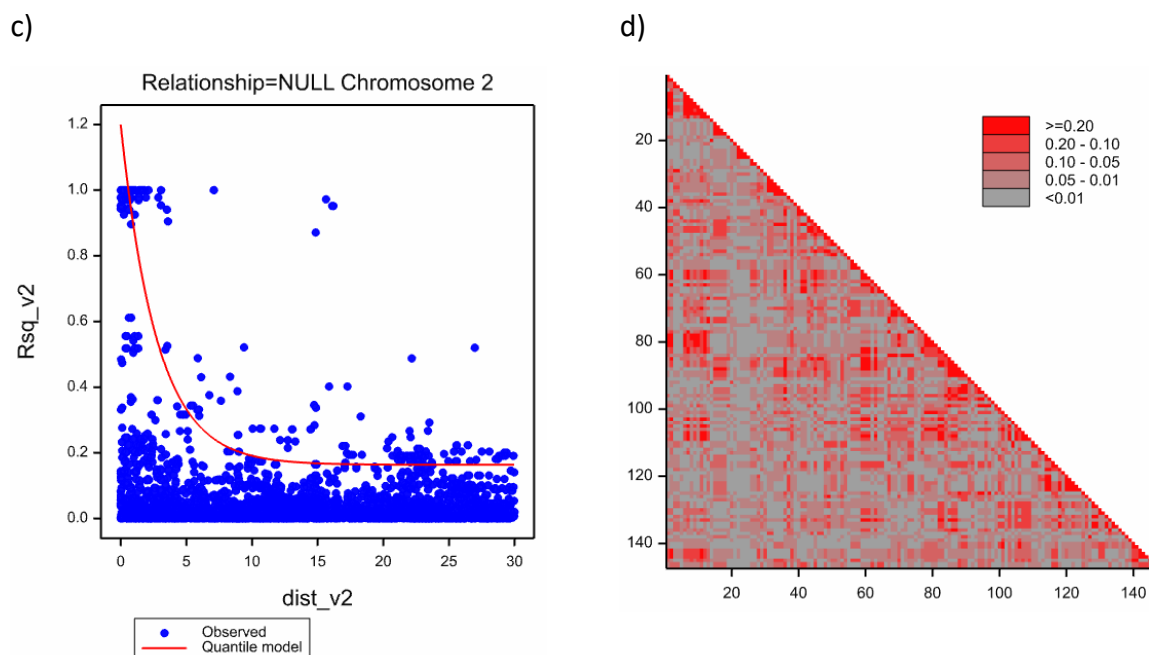
Figure 9.5: Graphical output from LD analyses of the MABDE barley chromosome 2 data: a) and b) are decay and shade plots, respectively, from the analysis with genetic relatedness modelled by *a priori* subpopulations, c) and d) are decay and shade plots, respectively, from the analysis without any correction for genetic relatedness.

9.4 Marker-trait association analysis

In Section 9.3, we demonstrated that LD between markers can be inflated by genetic relatedness. Similarly, a statistical association between a marker and a QTL can be the consequence of genetic relatedness. Therefore, models used to test for marker-trait associations must correct for genetic relatedness. Within Genstat, association mapping accommodates genetic relatedness by including in the mixed model either:

1) a kinship matrix with coefficients of co-ancestry between genotypes (Yu *et al.,* 2006; Malosetti *et al.*, 2007)

2) scores of the significant principal components (Section 9.2) as covariables (Price *et al*., 2006; Patterson *et al*., 2006)

3) a factor indicating subpopulation membership for each genotype (Zhao *et al*. 2007; Pasam *et al*. 2012)

It is also possible to run an analysis without correction for genetic relatedness (i.e. a naïve analysis). Such an analysis is generally only useful for purposes of comparison. We

describe the underlying statistical model for each strategy in turn, before illustrating marker-trait association analysis using the MABDE barley panel data.

Genstat performs association mapping in the mixed model framework, fitting markers as fixed and genotypes as random using REML (Malosetti *et al.*, 2007). Trait means (y_i) are used as input. If the raw plot (unit) data are available, read Chapter 3 to obtain trait means.

9.4.1 The null (naïve) model

Under the assumption of homogeneous genetic relatedness, the mixed model for marker-trait association analysis on a single-environment data set with bi-allelic markers can be expressed as:

$$y_i = \mu + \alpha x_i + (G_i + \varepsilon_i), \qquad\qquad \text{Equation 1}$$

where

y_i is the trait mean for genotype i

μ is the overall mean

α is the marker effect at the position being tested

x_i is the genetic predictor for genotype i at the marker position being tested (for bi-allelic markers $x_i \in (0,1,2)$)

G_i is the genotype i effect

ε_i is the error for genotype i.

The parentheses around the error, ε_i, and the random genotype effect, G_i, denote that these two terms cannot be separately estimated. Their joint effect ($G_i + \varepsilon_i$), the residual, is assumed to arise from a Normal distribution with mean 0 and variance σ^2. This imposes a genetic variance-covariance structure, VCOV(G_i), that assumes no genetic relatedness (or population structure) between genotypes. That is, for n genotypes

$$\text{VCOV}(\boldsymbol{G_i}) = \mathbf{I}_n \sigma^2,$$

an $n{\times}n$ matrix with σ^2 on the diagonals and zeroes on the off-diagonals.

9.4.2 Kinship model

The assumption of no genetic relatedness is unrealistic for the majority of association mapping panels, as most cultivar collections share some degree of relatedness. The kinship model provides an alternative parameterization for VCOV($\boldsymbol{G_i}$), with off-diagonal values determined by the degree of relatedness between the genotypes.

The kinship matrix, **K**, is a symmetric $n \times n$ matrix with 1s on the diagonal and co-ancestry coefficients ($0 \leq \theta \leq 1$) between all pairs of genotypes elsewhere. Assuming an additive genetic model, the genetic covariance between genotypes i and i^* with coefficient of co-ancestry θ_{ii^*} is: $COV(i, i^*) = 2\theta_{ii^*}\sigma_g^2$ (see Lynch and Walsh, 1998). Therefore, in the kinship model the genetic variance-covariance matrix takes the form:

$$VCOV(\boldsymbol{G}_i) = 2\mathbf{K}\sigma_g^2.$$

9.4.2.1　Forming a kinship matrix in Genstat

The Form Kinship Matrix menu can be used to construct a kinship matrix from marker scores. The coefficients of co-ancestries are calculated using either the Dice similarity measure or by simple Correlation. The menu is accessed from Stats | QTLs (Linkage/Association) | Genotypic Analysis | Form Kinship Matrix; or, in the QTL Data View via the shortcut Genotypic analysis | Form Kinship Matrix.

9.4.3　Eigenanalysis model

An alternative approach to modelling the genetic variance-covariance matrix, VCOV($\boldsymbol{G}_i$), is to use molecular marker data to group like genotypes. Eigenanalysis, where the resulting principal component scores represent population structure, is one such method (see Section 9.2). The eigenanalysis method, which effectively approximates the structuring of VCOV($\boldsymbol{G}_i$) by the kinship matrix, is less computationally intensive than the kinship model (Section 9.4.2).

The eigenanalysis model accommodates population structure by including as covariables the significant principal component scores (Patterson *et al*., 2006) in the mixed model. The mixed model can be expressed as:

$$y_i = \mu + \alpha x_i + \sum_{d=1}^{D} \alpha_d^* P_{i,d} \quad + (G_i + \varepsilon_i), \qquad \text{Equation 2}$$

where
$P_{i,d}$　　is the dth principal component score for genotype i
α_d^*　　the effect associated with the dth principal component
D　　is the number of significant principal components ($d = 1, …, D$).
All other terms are as described in Equation 1 (Section 9.4.1).

The effects associated with the D principal components ($\alpha_1^*, …, \alpha_D^*$) can be modelled either as fixed terms or random terms (with variance σ_d^2).

9.4.4 Subpopulation model

The mixed model platform can also incorporate population structure by stratifying the population using a previously determined subpopulation grouping factor. Correlation structure on the genotypes can be imposed by including the subpopulation factor as a random (or fixed) effect in the mixed model:

$$y_{i(k)} = \mu + \alpha x_{i(k)} + S_k + \left(G_{i(k)} + \varepsilon_{i(k)} \right), \qquad \text{Equation 3}$$

where

S_k is effect of the kth subpopulation ($k = 1, …, K$)

$i(k)$ denotes that genotype i is nested within subpopulation k.

All other terms are as described in Equation 1 (Section 9.4.1).

The subpopulation effects ($S_1, …, S_K$), modelled as a random with variance σ_k^2, impose equal correlations between genotypes within the same subpopulation, whilst genotypes in different subpopulations are assumed uncorrelated. Alternatively, population structure can be accommodated by fitting the subpopulation effects as fixed.

Subpopulations may derive from program STRUCTURE, which uses molecular marker information within a Bayesian framework to infer population structure (http://pritchardlab.stanford.edu/structure.html), or be based on, for example, geographic origin.

9.4.5 Association analysis in Genstat

Genstat performs genome-wide marker-trait association scans by testing the significance of the marker effect (α, a proxy for the QTL effect) using a marginal Wald test (Section 10.6; Searle *et al*., 1992; Verbeke and Molenberghs, 2000) at each marker location.

The Single Trait Association Analysis window (Figure 9.6) can be accessed from:

- Stats | QTLs (Linkage/Association) | QTL Analysis | Single Trait Association Analysis (Single Environment); or,
- in the QTL Data View via the shortcut QTL analysis | Single Trait Association Analysis (Single Environment).
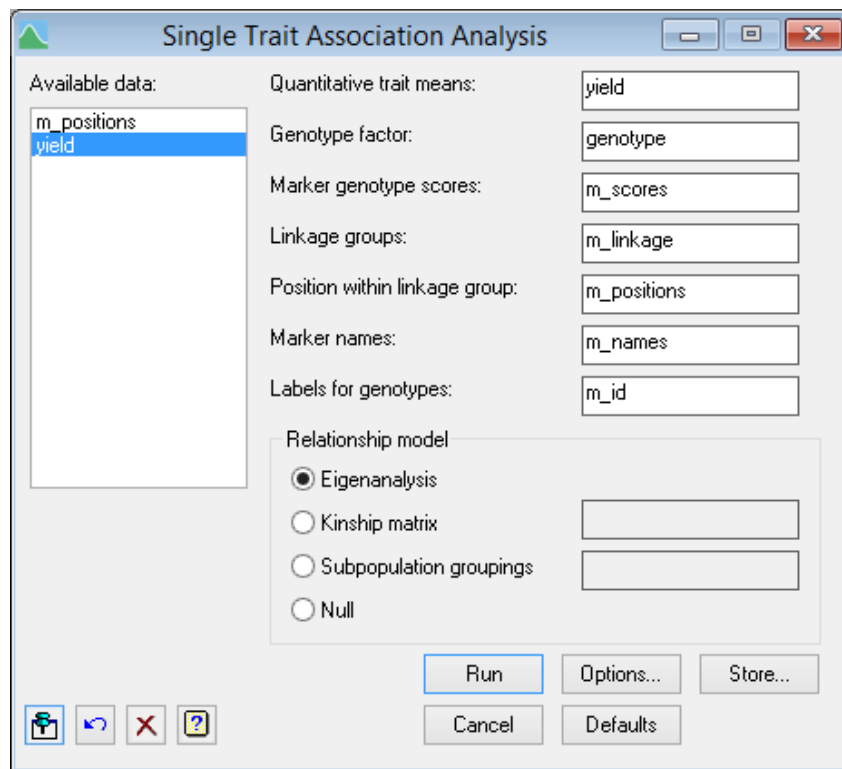
Figure 9.6: Single Trait Association Analysis window.

The trait for analysis, in this case `yield`, is specified in the Quantitative trait means: field. The fields for Genotype factor:, Marker genotype scores:, Linkage groups:, Position within linkage group:, Marker names:, and Labels for genotypes: will be automatically filled using data in the QTL Data Space. The Relationship model, accounting for genetic relatedness between genotypes, defaults to Eigenanalysis, but you can change this to the model of your choice. Selecting Kinship matrix requires the coefficient of co-ancestry matrix, here `m_kinship`, to be specified (Figure 9.7a), whereas selecting Subpopulation grouping requires the subpopulation grouping factor, here `group`, to be specified (Figure 9.7b).
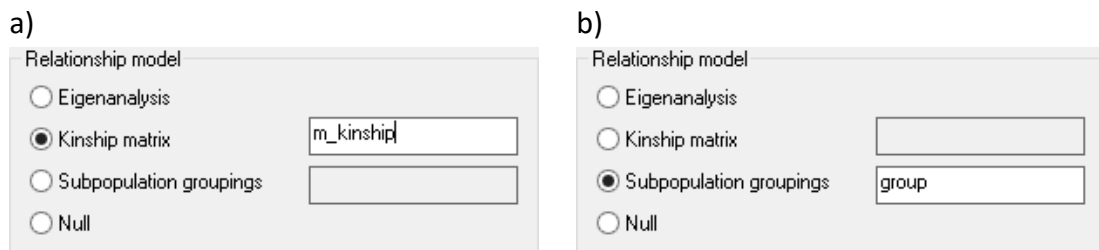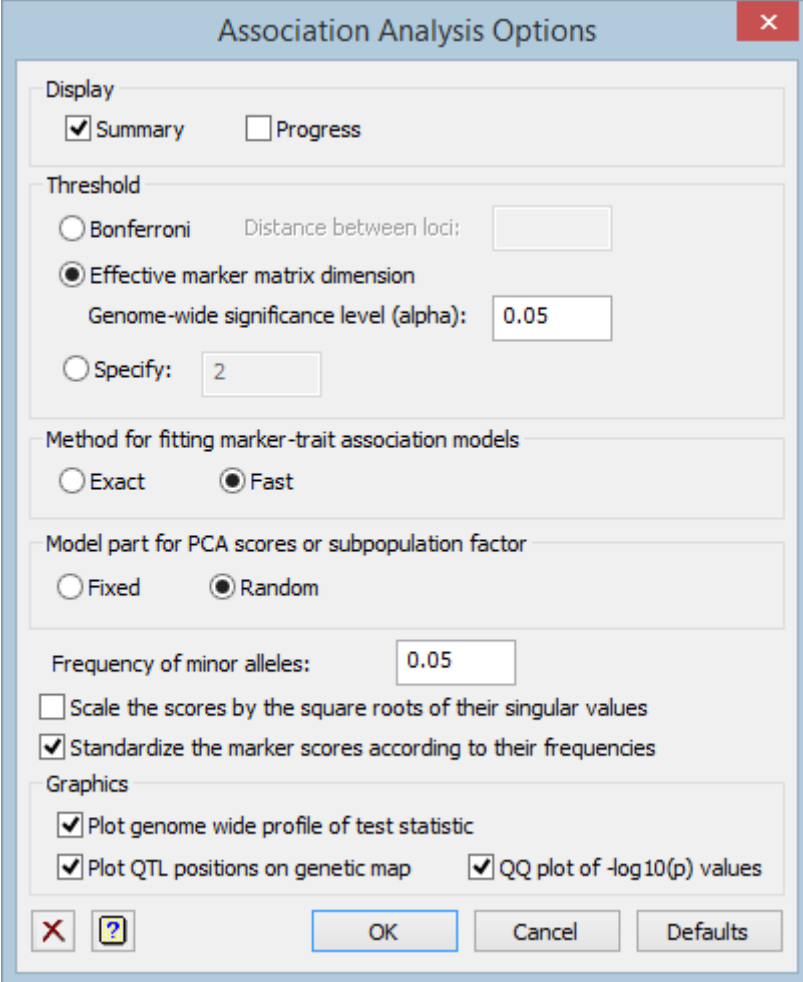
Figure 9.7: Specifying the model to account for genetic relatedness between genotypes. a) kinship model and b) subpopulation model.

Click the Options button to open the Association Analysis Options window (Figure 9.8).

Figure 9.8: Association Analysis Options window for a single environment data set.

The output produced is controlled by the Display and Graphics panes.

As association mapping involves conducting multiple significance tests along the genome, the Threshold box is used to specify a method to adjust for multiple comparisons. The Bonferroni method calculates the effective number of tests, assuming independent tests occur at a fixed distance along the genome. This distance (in cM) is specified in the Distance between loci: field. A study of LD decay (see Section 9.3) helps inform the appropriate distance, i.e. the distance at which LD is no longer high (in this example, ~5cM). If the distance is not set, an independent test is assumed at every marker, a very

conservative assumption in most cases. Alternatively, Effective marker matrix dimension determines the effective number of columns (nC) in the marker matrix data, using the estimator given by Patterson *et al.* (2006), and calculates the threshold as -$\log_{10}\left(\frac{alpha}{nC}\right)$. The parameter *alpha* is the genome-wide Type I error rate, which is specified in the field Genome-wide significance level (alpha): (default 0.05). Finally, a user-defined threshold (on the -$\log_{10}$ scale) can be set in the Specify: field. The default 2 is equivalent to *alpha* = 0.01.

The pane Method for fitting marker-trait association models provides two model fitting options; Exact or Fast. The Exact method solves the mixed model for each marker separately. Conversely under the Fast method, the mixed model is only solved for the genetic background model, i.e. the model without markers. The estimated variance-covariance matrix from the genetic background model is then used to perform a generalized least squares scan for all the markers. Note, the Fast method can only be implemented for bi-allelic markers.

When the Exact method is used, the principal components scores and the subpopulation factor (if the relationship model is set to Eigenanalysis or Subpopulation groupings, respectively) can be included as either Random or Fixed terms in the mixed model. The Model part for PCA scores or subpopulation factor box controls this setting; default Random. Under the Fast method, they are always fitted as random.

The Frequency of minor alleles: field specifies the frequency *q* below which alleles are considered "*rare*" (default 0.05). For multi-allelic markers, rare alleles are pooled together. Markers whose most frequent allele occurs $\geq$ (1-q)$\times$100% of the time are considered close to fixation and not used in the analysis.

Scaling and standardizing of the genetic predictors is controlled using the options Scale the scores by the square roots of their singular values and Standardize the marker scores according to their frequencies, respectively.

The Store button opens the Association Analysis Store Options window (Figure 9.9) allowing you to save results from the analysis (i.e. marginal Wald statistics and associated -$\log_{10}$(p-values), and information and results from the significant markers). Names for the saved data structures need to be specified in the corresponding In: fields. Checking Display in Spreadsheet displays the saved results within a new spreadsheet.
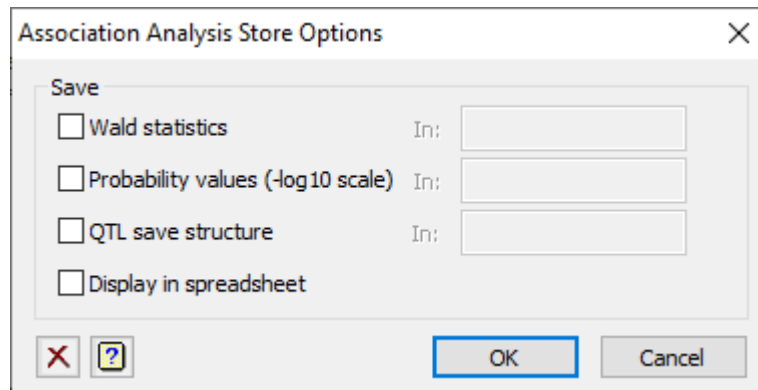
Figure 9.9: Association Analysis Store Options window.

We perform an association mapping analysis, using Eigenanalysis, on the MABDE barley data (Section 1.3.4) with the options shown in Figure 9.8. The first section of output provides some summary information on the MABDE data, and describes the model and options used in analysis.

```
Status: fitting mixed model without marker information.

Status: performing GWAS scan.

Summary of marker-trait association results
===========================================


Trait: yield
Number of markers: 811
Number of linkage groups: 7
Number of genotypes: 179

GWAS model: EIGENANALYSIS
Substructure covariable/factor: RANDOM
GWAS method: FAST
Threshold method: NEFFECTIVE
Number of assumed independent tests: 28.13

Threshold for significance: 2.75
Inflation factor (lambda): 1.363
CPU time for data preparation: 0:00:17
CPU time for GWAS scan: 0:00:00
```

The second section lists the markers with a significant association with the trait of interest. Here, two markers, one on chromosome 5 (D5042) and one on chromosome 7 (D7050) are significantly associated with yield.

```
List of significant marker-trait associations
=============================================
```

Index	Marker	Linkage group	Position	-log10(P)	Number of alleles
507	D5042	5	74.5	3.728	2
729	D7050	7	72.4	2.881	2

The final section of output provides estimates of the marker effects (from the significant marker-trait association). Values reported under the `Frequency`, `Effect` and `Std.error` columns correspond to the allele given in the `Allele` column. In this case, at marker `D5042` allele 1 occurs with frequency 0.33. The estimated effect at `D5042` is 0.2139 with standard error 0.0582. Therefore replacing allele 0 by allele 1 at `D5042` is expected to result in an increase yield of 0.2139 units.

Index	Marker	Reference allele	Allele	Frequency	Effect	Std. error
507	D5042	0	1	0.3296	0.2139	0.0582
729	D7050	1	0	0.3128	-0.2294	0.0721

The profile plot displays p-values (on the -log$_{10}$ scale) from marginal Wald tests of the marker effects (α) along the chromosomes (Figure 9.10). Each chromosome (or linkage group) is depicted by a different colour. The red horizontal line represents the threshold level, above which the null hypothesis of no marker effect is rejected. In this example, the Effective marker matrix dimension method was used (see Figure 9.8), resulting in a significance threshold of 2.75.
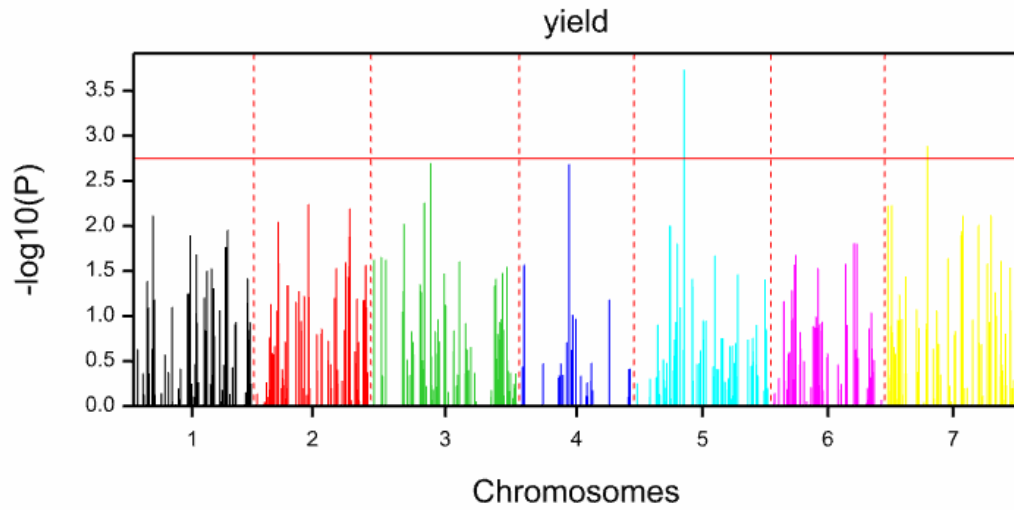
Figure 9.10: Profile plot from association mapping (using eigenanalysis) of `yield` for the MABDE barley data.

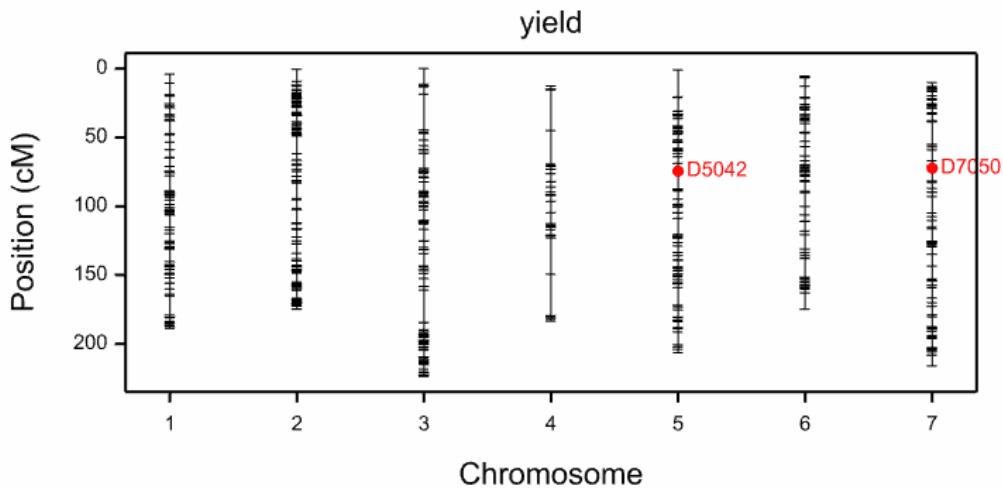In Figure 9.11 the significant markers ("QTLs") are plotted on a genetic map.

Figure 9.11: Genetic map of the MABDE barley data with the two detected QTLs shown.

Figure 9.12 gives the quantile-quantile (QQ) plot of the -$\log_{10}$(p-values) from marginal Wald tests. The QQ plot graphically represents the deviations of the observed p-values from the null hypothesis of no association. Deviations from the diagonal line suggest that either the null hypothesis is incorrect, population structure (i.e. genetic relatedness) has not been corrected for, or the presence of a very highly associated region containing many markers. The effect of not correcting for population structure is clearly demonstrated in

the QQ plot from the naïve analysis (Figure 9.13), i.e. where the relationship model is set to Null. Here, the large deviations are indicative of spurious associations. Note, a sharp and sudden series of deviations would indicate the presence of a very highly associated region containing many markers.
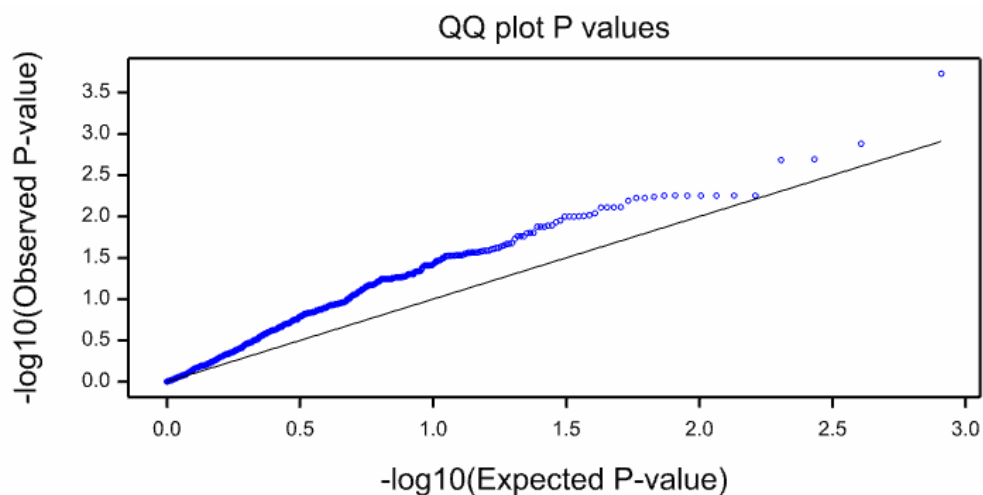

Figure 9.12: QQ plot from association mapping (using eigenanalysis) of the MABDE barley data.
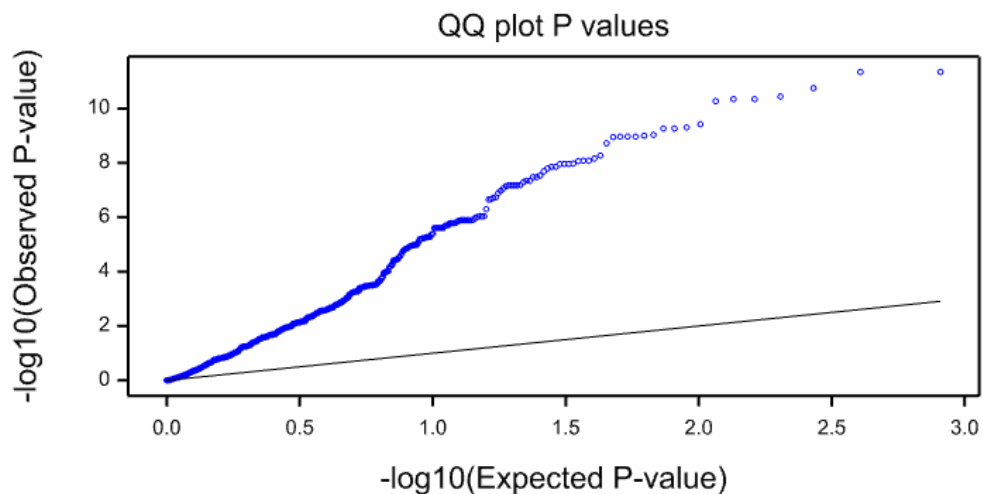

Figure 9.13: QQ plot from association mapping of the MABDE barley data where no correction is made for population structure.

9.5 Multi-environment marker-trait association mapping

Genstat also allows association mapping in multiple environments. The menus and statistical methodology are similar to the single environment case. However, now the genome-wide scan tests for both marker main effects (a proxy for QTL main effects) and marker by environment interactions (a proxy for QTL×E interactions). Also, importantly, an appropriate variance-covariance model for describing the variation between genotypes both across and within environments must be specified.

The Single Trait Association Analysis (Multiple Enviro window (Figure 9.14) can be accessed from either via Stats | QTLs (Linkage/Association) | QTL Analysis | Single Trait Association Analysis (Multiple Environment); or, in the QTL Data View via the shortcut QTL analysis | Single Trait Association Analysis (Multiple Environment).

Genstat will automatically populate the input fields using data from the QTL Data Space. The menu is similar to that of the single environment analysis (Figure 9.6) except for the addition of the Environment Factor: and Variance-covariance model: fields. Also note that the kinship method to account for genetic relatedness (Section 9.4.2) is not available in the multiple environment context.
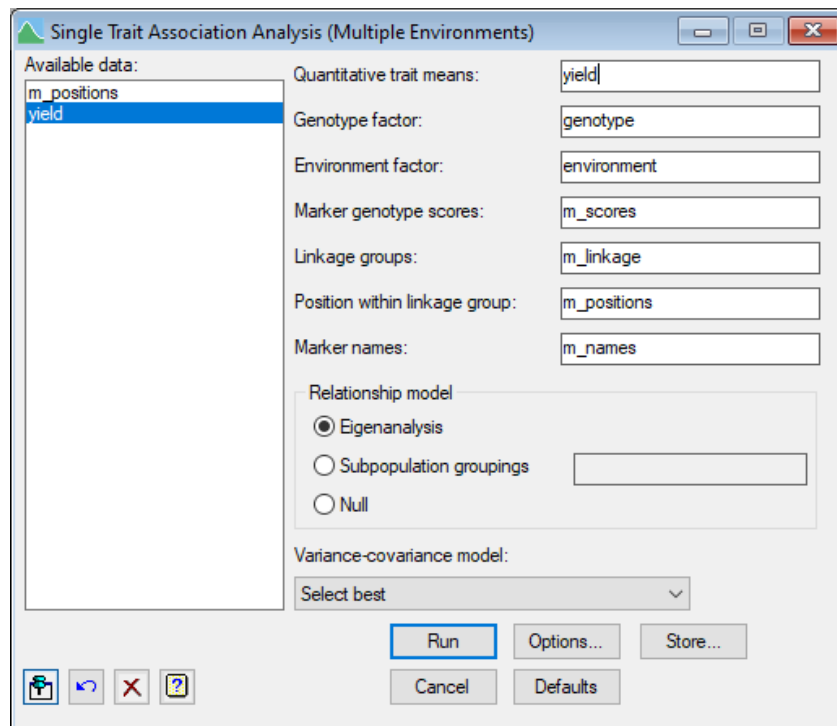

Figure 9.14: Menu window for multiple environment marker-trait association analysis.

In the Variance-covariance model: field the variance-covariance structure for modelling the variation between genotypes both across and within environments is selected (see Section 4.1.3). If you've previously performed a genotype by environment analysis (as described in Chapter 4), you can select this model here. If not, the Select best setting will automatically select the best variance-covariance model according to the criterion set in the Association Analysis Options window (Figure 9.15): either Schwarz information criterion (SIC) or Akaike information criterion (AIC).

In the Association Analysis Options window you can also specify what output to display, the threshold value above which the null hypothesis of no marker effect is rejected, whether to fit the PCA scores or subpopulation factor as fixed or random, the frequency below which alleles are considered rare, how to scale and standardize the genetic predictors, and which graphics to produce. The default threshold is 2, equivalent to *alpha* = 0.01. However, you may find this too liberal and opt to increase the threshold level to decrease the probability of detecting non-significant markers.
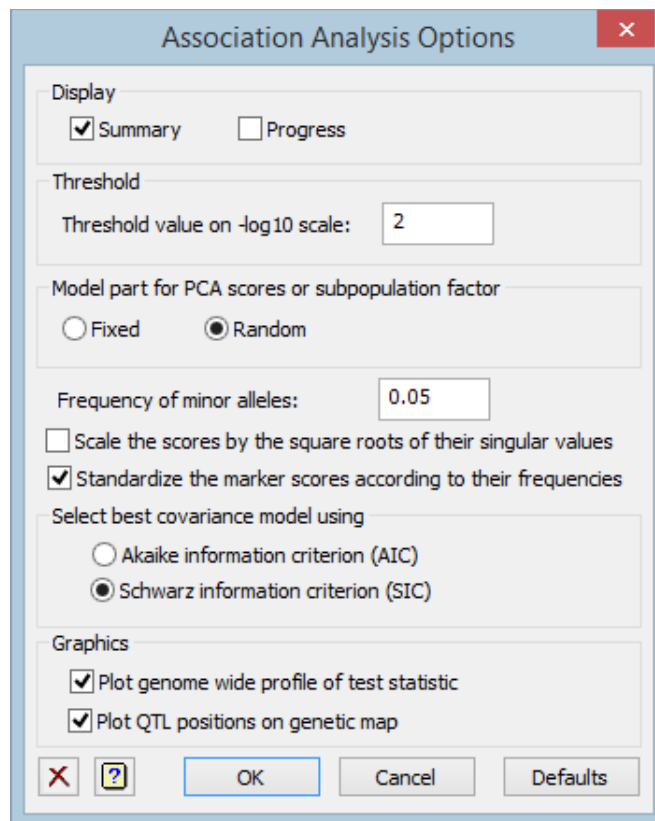

Figure 9.15: Association Analysis Options window for a multiple environment data set.

The genome-wide scan comprises 2 steps:
1) a Wald test is used for each marker, individually, to test the null hypothesis that the marker effect is zero in every environment
2) if the null hypothesis is rejected, a second test is performed to check whether the marker-by-environment interaction is significant.

Full documentation is provided within the Genstat Help System (🔳). Refer to procedure QMASSOCIATION.

9.6   Multi-allelic markers

In the previous sections, association mapping was described for bi-allelic markers, such as SNPs. However, Genstat's facilities can also accommodate multi-allelic markers. Here, the genetic predictor takes the form of a matrix of allele frequencies instead of a variate. The most frequent allele is set as the reference level. Eigenanalysis (Section 9.2), LD analysis (Section 9.3), and association mapping (Section 9.4) then proceed analogously to that previously described.

For example, consider a marker with three alleles: A, B, and C. Setting allele A as the "*reference*" allele, the genetic predictor matrix for genotypes AA, BB, CC, AB, AC, BC (in this order) is:

$$\begin{bmatrix} 0 & 0 \\ 2 & 0 \\ 0 & 2 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}',$$

where the first column corresponds to the number of B alleles, and the second to the number of C alleles.

9.7   References

Flint-Garcia, S.A., Thornsberry, J.M., & Buckler, E.S. (2003). Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology*, **54**, 357-374.

Kraakman, A.T., Nik, R.E., Van den Berg, P.M., Stam, P., & van Eeuwijk, F.A. (2004). Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics*, **168**, 435-446.

Lynch, M., & Walsh, B. (1998). Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, MA.

Malosetti, M., van der Linden, C.G., Vosman, B., & van Eeuwijk, F.A. (2007). A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. *Genetics*, **175**, 879-889.

Patterson, N., Price, A.L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, **2**, 2074-2093.

Pasam, R.K., Sharma, R., Malosetti, M., van Eeuwijk, F.A., Haseneyer, G., Kilian, B., & Graner, A. (2012). Genome-wide association studies for agronomical traits in a worldwide spring barley collection. *BMC Plant Biology*, 12:16.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904-909.

Pritchard, J.K., Stephens, M., & Donnelly, P. (2000a). Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945-959.

Pritchard, J.K., Stephens, M., Rosenberg, N.A., & Donnelly, P. (2000b). Association mapping in population structure. *American Journal of Human Genetics*, **67**, 170-181.

Searle, S.R., Casella, G., & McCulloch, C.E. (1992). Variance components. Wiley, New York.

Verbeke, G., & Molenberghs, G. (2000). Linear mixed models for longitudinal data. Springer-Verlag Inc.: Berlin; New York.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S., & Buckler, E.S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, **38**, 203-208.

Zhao, J., Paulo, M.J., Jamar, D., Lou, P., van Eeuwijk, F., Bonnema, G., Vreugdenhil, D., & Koornneef, M. (2007). Association mapping of leaf traits, flowering time, and phytate content in *Brassica rapa*. *Genome*, **50**, 963-973.

Zhu, C., Gore, M., Buckler, E.S., & Yu, J. (2008). Status and prospects of association mapping in plants. *Plant Genome*, **1**, 5-20.

10 Introduction to linear mixed models

In this chapter, we provide a brief introduction to linear mixed models and their implementation in Genstat. Within the QTL system, linear mixed models are used for both preliminary analysis and QTL detection, and this chapter can be used as a reference for the underlying methods and models. It may be skipped on a first reading. More detailed information on linear mixed models and REML estimation can be found in *A Guide to REML in Genstat*. Here, we focus on the aspects most relevant to the QTL system, using analysis of a simple field trial as an example.

In this chapter you will learn:

- the form of a linear mixed model
- the distinction between fixed and random terms
- how to interpret the random model
- how to compare the fit of different random models
- how to assess the importance of fixed model terms
- how to check the model assumptions
- a recipe for identification of a suitable linear mixed model for a designed experiment

10.1 The linear mixed model

A linear mixed model is specified using two components, known as the fixed and random models. In general, the choice of which terms to classify as fixed and which as random may depend on the aims of the analysis. From the perspective of designed experiments, terms representing experimental treatments are usually assigned as fixed, and terms associated with the randomization structure of the design are assigned as random. In general, fixed terms often represent the effect of specific conditions applied or chosen for the experiment, i.e. the experimental treatments. Random terms often represent terms where the conditions observed comprise a sample from some wider population, and it is the variability of the population that is of interest. The structural (or randomized) components of an experimental design, such as blocks and plots, can usually be argued to fall into this category.

  As an example, we use the CIMMYT spring wheat trials described in Section 1.3.3. These trials tested 169 lines of spring wheat using two replicates of a lattice design. We will consider the data from trial `HEAT05`, data contained in Genstat spreadsheet file `SB_HEAT05.gwb`. In a standard analysis of this trial, the 169 lines (factor `Genotype`) would be considered as a set of fixed effects. The design consists of two replicates (factor `Rep`), each containing 13 sub-blocks (factor `Subblock`) with 13 plots. This gives a nested blocking structure, written as `Rep/Subblock`. The two components of the model can therefore be written as

Fixed model: `Genotype`
Random model: `Rep/Subblock`   (Model 1)

  Occasionally, other arguments are used to assign terms as random rather than fixed. One reason for this is that predicted random effects can be more precise than fixed effects - this is explained further in Section 10.5 below. If precision is the most important criteria for a prediction, then it may be preferable to assign terms as random. This argument is often used in plant breeding trials, where genotypes may be assigned as random in order to increase precision and avoid selection bias. In that case, the two components of the model are written as

Fixed model:
Random model: `Genotype + Rep/Subblock`   (Model 2)

Note that a constant term will automatically be added to the fixed model by default. The response variate (called the Y-variate in Genstat) is required to complete specification of the model.

In a simple model, the effects associated with each random term and the residual term are assumed to be a set of independent samples from a Normal distribution. Effects within each term have a common variance, which is known as the variance component for that term. In addition, it is assumed that effects from different random terms are independent.

For the HEAT05 trial, Model 1 (genotypes fixed, design factors random) can be written in terms of the individual observations as

$$y_{ijk} = \mu + G_{g(ijk)} + R_i + B_{ij} + e_{ijk} \qquad \text{(Model 1)}$$

where

- there are 338 observations, labelled by the replicate ($i = 1,2$), sub-block within replicate ($j = 1, \ldots, 13$) and plot within sub-block ($k = 1, \ldots, 13$)
- y_{ijk} is the observed response on the kth plot within the jth sub-block within the ith replicate
- μ is a constant (or intercept) term
- G_g is the effect of the gth genotype
- $g(ijk)$ indicates the genotype randomly allocated to the ijk^{th} plot
- R_i is the random effect associated with the ith replicate with variance component σ_R^2
- B_{ij} is the random effect associated with the jth sub-block in the ith replicate with variance component σ_B^2
- e_{ijk} is the random deviation for the kth plot within the jth sub-block within the ith replicate, with residual variance σ^2.

There are two fixed terms here: the constant (μ) and the set of genotype effects (G_g, $g = 1, \ldots, 169$). The fixed model uses first-level-zero parameterization (see Genstat Statistics Guide, Section 5.2.2, for further details) so μ estimates the predicted mean for the first genotype, G1 is constrained equal to zero, and G_n represents the effect of genotype n as a deviation from the first genotype. There are two random terms: the set of 2 replicate effects (R_i) and the set of 26 replicate by sub-block effects (B_{ij}), plus the residual term (deviations).

This model can be fitted using the Stats | Mixed Models (REML) | Linear Mixed Models menu (Figure 10.1 and Figure 10.2).
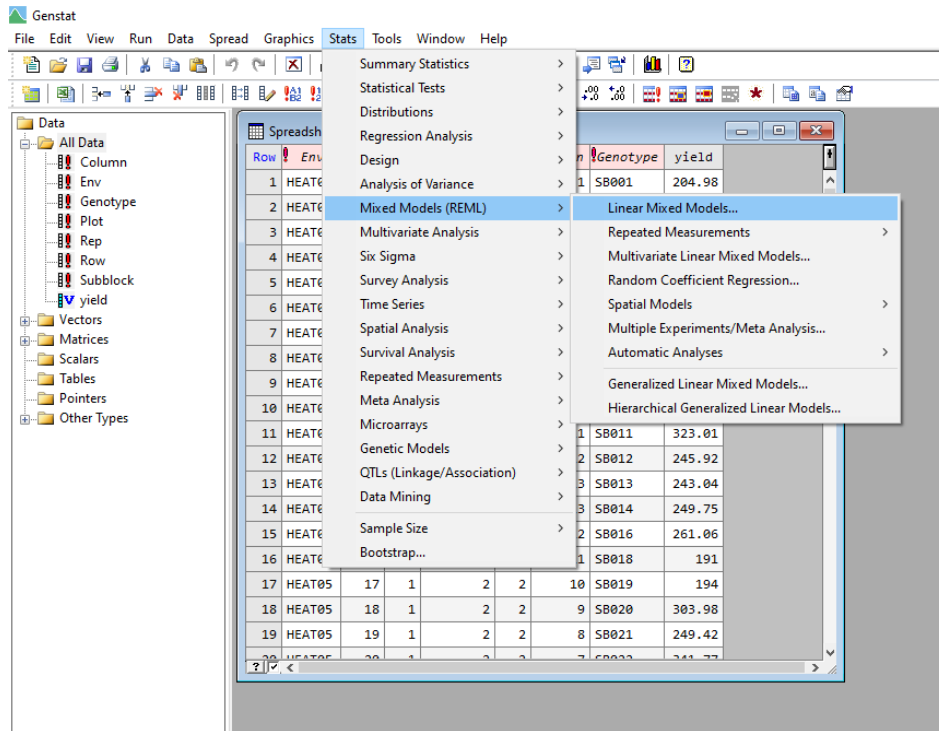
Figure 10.1: Accessing the Linear Mixed Models menu.
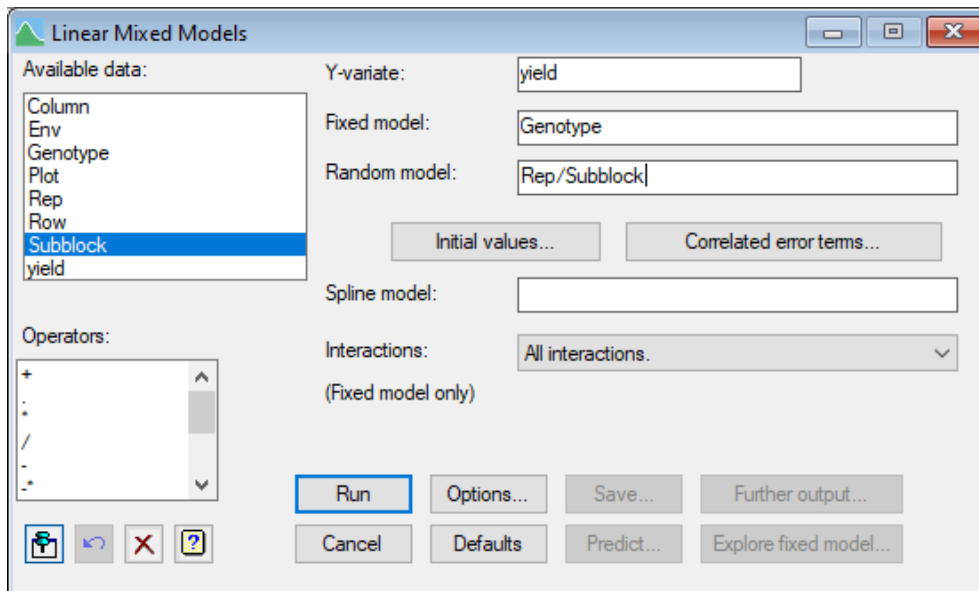

Figure 10.2: The Linear Mixed Models menu.

If we choose to display the Model and Variance Components, using the Options button, then we obtain the following output:

```
REML variance components analysis
=================================


Response variate:  yield
Fixed model:       Constant + Genotype
Random model:      Rep + Rep.Subblock
Number of units:   338

Residual term has been added to model

Sparse algorithm with AI optimisation


Estimated variance components
-----------------------------


Random term              component      s.e.
Rep                         1215.9     1742.7
Rep.Subblock                 179.3       71.2


Residual variance model
-----------------------


Term             Model(order)  Parameter      Estimate      s.e.
Residual         Identity      Sigma2            428.7      50.5
```

The output starts with the model summary. The constant is automatically added into the model, and so there are two fixed terms, `Constant` and `Genotype`. There are two random terms, `Rep` and `Rep.Subblock`, and a residual term has been automatically added into the model. The variance components for the random terms are listed together. The variance component for the `Rep` term, i.e. the variance of the replicate effects (R_i), is estimated as 1215.9, and the variance component for the `Rep.Subblock` term, i.e. the variance of the sub-block effects (B_{ij}), is estimated as 179.3. The residual variance is listed separately, and is estimated as 428.7. Interpretation of these values is considered in Section 10.2.

The estimated parameters of the linear mixed model are the set of fixed effects and the variance parameters. The random effects have a slightly different status, which is discussed further in Section 10.5. Variance parameters are estimated by REML (residual maximum likelihood, also called restricted maximum likelihood), a method introduced by Patterson & Thompson (1971). The fixed effects are then estimated by the method of

212

generalized least squares, conditional on the estimated values of the variance components. One advantage of the REML method is that it gives the same estimates of fixed effects and their standard error of the differences (SEDs) as obtained from multi-stratum ANOVA when the structure is balanced and, where treatment information is divided across strata, estimates will be combined efficiently across strata into a single estimate (see *A Guide to REML in Genstat* for further details).

10.2 Understanding the random model

In a simple random model, with uncorrelated effects, the variance components associated with the random terms generate a variance-covariance matrix for the observations. The variance of an observation is equal to the sum of the variance components and it can be derived from the algebraic form of the model: the variance of the fixed effects is zero, and the variance of each random effect is equal to its variance component. The covariance between any two observations depends on the number of random effects held in common across the observations, and is the sum of the variance components for these common random effects.

For the `HEAT05` field trial, with `Genotype` fitted a fixed effect and `Rep` and `Rep.Subblock` effects fitted as random terms, the total variance of an observation is just the sum of the variance components:

$$
\begin{aligned}
\text{var}(y_{ijk}) &= \text{var}(R_i) + \text{var}(B_{ij}) + \text{var}(e_{ijk}) \\
&= \sigma_R^2 + \sigma_B^2 + \sigma^2 \\
&= 1215.9 + 179.3 + 428.7 \\
&= 1823.9
\end{aligned}
$$

It follows that omitting a random term from the model is equivalent to setting its variance component equal to zero. The covariance between yields from two plots in the same incomplete block is then the sum of the replicate and sub-block variance components, derived as

$$
\begin{aligned}
\text{cov}(y_{ijk}, y_{ijl}) &= \text{cov}(R_i + B_{ij} + e_{ijk}, R_i + B_{ij} + e_{ijl}) \\
&= \text{var}(R_i) + \text{var}(B_{ij}) \\
&= \sigma_R^2 + \sigma_B^2 \\
&= 1215.9 + 179.3 \\
&= 1395.2
\end{aligned}
$$

Similarly, the covariance between yields from two plots in the same replicate is equal to the replicate variance component, i.e. 1215.9, and the covariance between two plots in different replicates is zero, as they have no random terms in common. The variance model generated by the design thus implies that measurements from plots within the same incomplete block are slightly more highly correlated than those within the same replicate but in different sub-blocks, and that there is no correlation between measurements from separate replicates. Other patterns of covariance can be generated by using different random terms or correlated random effects.

Whilst the definition of the variance components as variances of the random effects requires that the variance components are positive, in fact the variance structure requires only that the total variance is positive and, more generally, that the variance of any linear combination of observations is also positive (this property is known as positive-definiteness). In general, as we use random terms to reflect structure, we expect that units with random effects in common will be more similar than units without, and so generally variance components are expected to be positive. However, in some circumstances it is natural to allow variance components to be negative. For example, in field experiments, blocks are laid out on areas of ground thought to be reasonably homogeneous with respect to fertility and other trends. If a mistake is made, so that blocks are laid out in the wrong direction, then plots within the same block may be less alike than plots in different blocks, and this can only be modelled by using a negative variance component for blocks. Even if variance components are positive, it is possible that they may be estimated as negative values, due to sampling variability. For example, a variance component with true value equal to zero has a 50% chance of being estimated as a negative value. Genstat allows the user to specify whether estimates should be constrained to remain positive, via the Initial Values button on the Linear Mixed Models menu (Figure 10.2).

Estimates of variance components are presented with their standard errors (SEs). In many cases, some of the estimated variance components are small compared to their SEs and it might be natural to think of dropping these terms from the model. We advise against this course for two distinct reasons. Firstly, the SEs for variance components are only reliable for testing when there is a large amount of information contributing to the estimate; again, the SEs depend on an asymptotic approximation. A better approach is the use of **likelihood ratio tests** (LRTs), these are described in Section 10.4 below. Secondly, some of the random terms will have been included to describe the randomization structure of the design. This structure is a property of the design and is used to get appropriate degrees of freedom (df) for approximate F tests: the omission of some terms means that the random model is no longer serving this purpose.

10.3   More complex random models

So far, we have assumed that effects within a single random term are independent with a Normal distribution with common variance. We retain the assumption of a Normal distribution, but can relax the assumptions of independence and common variance, leading to a much wider range of variance models. The full range of variance models provided in Genstat is given in the *Statistics Guide* (Section 5.4 of *The Guide to the Genstat Command Language, Part 2 Statistics*), with some examples of their use given in *A Guide to REML in Genstat*. In this Guide, we focus on spatial models used within the QTL system for analysis of a single field trial (Section 3.6) and covariance modelling for genotype by environment (G×E) interactions (Chapter 4).

10.4   Comparison of random models: likelihood ratio tests

When using REML estimation, the log-residual likelihood value (ℓ_R) can be used to compare nested models that have different random terms but the same set of fixed terms (Welham & Thompson, 1997). In Genstat, the log residual likelihood is obtained via the deviance, denoted *D*, which is calculated by omitting some constant terms (with respect to the variance parameters) from the log-residual likelihood, then multiplying by -2. A random model that fits well will have a relatively high value of ℓ_R and hence a relatively low value of the deviance but, because of the omission of the constant terms, the absolute scale of the deviance is arbitrary and it may even be negative. The deviance for a term can be viewed using Options or Further Output then check the Deviance box under Display. The deviance for the standard model for the `HEAT05` trial is shown below:

```
Deviance: -2*Log-Likelihood
---------------------------

                Deviance   d.f.
                 1348.11   166

Note: deviance omits constants which depend on fixed model fitted.
```

The message printed below the deviance highlights that fact that one of the omitted constants is related to the fixed model. The presence of this constant in ℓ_R is the reason why it cannot be used to compare models with different fixed terms, and this is the case whether or not it is omitted from calculation of the log-residual likelihood.

Two nested random models (with the same fixed terms) can be compared using the difference between their deviances. We denote the larger model as M_0 with deviance D_0, and drop one or more random terms to obtain the smaller model, denoted as M_1 with deviance D_1. We wish to test the null hypothesis that the larger model gives no improvement over the smaller model. In the context of a variance components model where we wish to drop one term, this implies that the variance component of the dropped term is equal to zero. If the variance parameters in the dropped terms are unconstrained then under the null hypothesis the change in deviance, defined as D_1-D_0, is ≥ 0 is asymptotically distributed as a chi-square variable with df equal to the number of variance parameters in the dropped terms. For example, if the smaller model is constructed by dropping a single variance component from the random model, then the change in deviance has a chi-square distribution on 1 df.

The situation becomes more complicated when one or more of the variance parameters is constrained, and a constraint boundary coincides with the reduced model. This is the case in a variance components model when the null hypothesis is that one of the variance components is zero and that variance component is constrained to stay positive. In this case, the asymptotic distribution of the change in deviance becomes a mixture of chi-square distributions (Stram & Lee, 1994). For the case of a single variance component, this becomes a 50:50 mixture of a chi-square distributions with 0 and 1 df. In practice, this adjustment requires the halving of p-values obtained with respect to the chi-square distribution on 1 df. In other cases, the mixtures and resulting calculations become more complex. Crainiceanu & Ruppert (2004) argued that the Stram & Lee (1994) results do not apply to many mixed models, and this is an area of ongoing research.

An alternative approach uses information criteria to compare models, which again must have the same fixed terms but may have different random terms. In this case, there is no requirement for the random models to be nested. The Akaike and Schwarz information criteria (AIC and SIC, respectively) are defined as:

$$AIC = -2\ell_R + 2t = D + 2t$$
$$SIC = -2\ell_R + t \log(n - p) = D + t \log(n - p)$$

where t is the number of variance parameters in the model, n is the number of observations, and p is the number of independent fixed parameters. SIC is also known as the Bayesian Information Criterion (BIC). Both AIC and SIC are based on the deviance plus a penalty based on the number of variance parameters estimated and, in the case of SIC, the number of observations used in the log-residual likelihood (or the number of residual contrasts, in the terminology of Patterson & Thompson (1971)). The penalty is intended to offset the improvement in likelihood against the lack of parsimony implied

by adding more parameters into the model. Using this definition, smaller values of AIC or SIC indicate a better fit of the variance model. The SIC tends to be more conservative than AIC, i.e. to prefer models with fewer variance parameters. The selection of models using these criteria is demonstrated in the context of spatial analysis of a single trial (Chapter 3) and modelling genotype by environment interactions (Chapter 4).

10.5 Predictors of random effects (BLUPs)

We stated earlier that the parameters of the linear mixed model are the fixed effects and the variance components, but we wrote down our models in terms of fixed and random effects. In this section, we resolve this apparent contradiction and discuss the status of the random effects.

The linear mixed can be written in two forms. The form above, written in terms of both the fixed and random effects, is known as the conditional model since the response depends (is conditional) on the random effects. The marginal form of the model is obtained by integrating over the population of random effects: this gives a model with the expected value of the observations determined by the fixed terms alone, and the variance-covariance matrix of the observations is that generated by the random terms (as described above). Estimation takes place on the marginal model, whose parameters are the fixed effects and the variance components. However, we are still often interested in the values of the random effects, and so would like to obtain an estimate of their values, although clearly this only makes sense when the associated variance component is positive (as otherwise the random effects cannot be defined).

Because the random effects are not true parameters we obtain predictors, rather than estimates, of their values. These predictors are often called BLUPs, which is an acronym for best linear unbiased predictors. In this context, the label 'unbiased' can be slightly misleading, as it means that the expected value of the predictor is equal to the expected value of the population, which is zero. This does not mean that the expected value of the predictor is equal to the true random effect, and in fact the predictors are biased towards zero, a property known as shrinkage. This is illustrated in Figure 10.3, where predicted means generated with the `Genotype` factor in the random model are plotted against predictions generated with the `Genotype` factor in the fixed model.
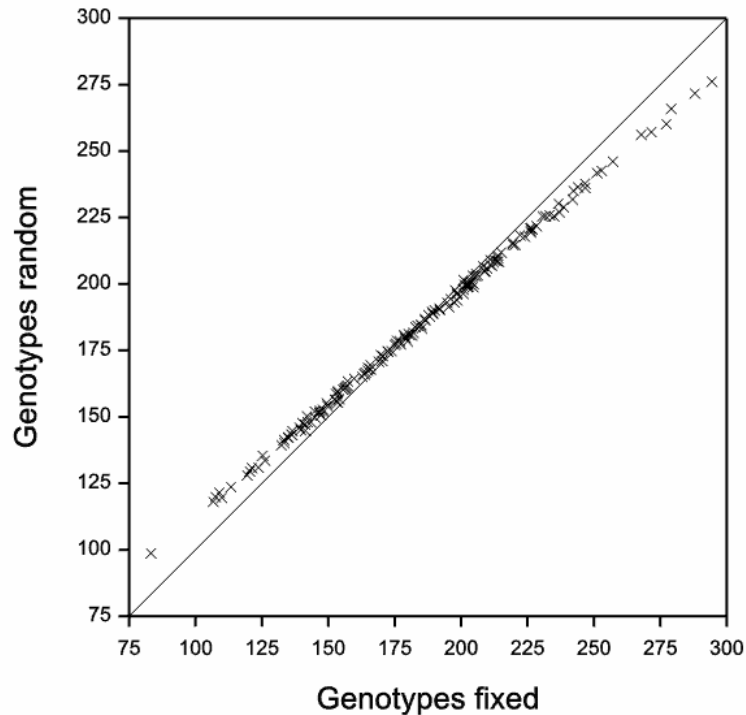
Figure 10.3: Genotype predictions (×) from `HEAT05` trial based on random or fixed genotype effects with 1:1 line.

The adjective 'best' means that these predictors have minimum mean squared error (defined as variance plus squared bias), conditional on the estimated variance components. This property of minimum mean squared error is attractive where accuracy in prediction is more important than unbiasedness, and is sometimes used as a justification for assigning terms to the random rather than fixed model, particularly in the context of variety evaluation (see Smith *et al.*, 2005, for discussion in this context).

10.6   Assessing fixed model terms

The estimates of the fixed effects used with REML are often called BLUEs, which is an acronym for best linear unbiased estimates. The property of unbiasedness means that the expected value of the estimator is equal to the true parameter value. In this context, 'best' means that these estimates have minimum variance within the class of unbiased estimators, conditional on the estimated variance components.

We usually wish to investigate the contribution of individual terms within the fixed model in explaining patterns of response and this is achieved using Wald tests. If the set

of fixed terms is non-orthogonal, it is necessary to consider both incremental and marginal forms of these statistics; incremental statistics reflect the change in fit on sequentially adding individual terms into the fixed model, and marginal statistics reflect the change on dropping individual terms from the full model. On dropping terms, we respect marginality, for example, we do not drop a term if it is involved in any higher-order interactions (for further information see the *ANOVA and Design Guide* or *The Guide to the Genstat Command Language, Part 2 Statistics*). If we use Options on the Linear Mixed Models menu to display Wald Tests, then we get the following output:

```
******** Warning 6, code VD 39, statement 1 on line 170

Command: REML [PRINT=waldTests; FMETHOD=automatic; MVINCLUDE=*; METHOD=AI;
MAXCYCLE=20] yield; SAVE=_remlsave
Error in AI algorithm when forming denominator DF for approximate F-tests.


Wald tests for fixed effects
----------------------------

Sequentially adding terms to fixed model

Fixed term                  Wald statistic   d.f.    Wald/d.f.  chi pr
Genotype                          1168.72    168         6.96  <0.001

Dropping individual terms from full fixed model

Fixed term                  Wald statistic   d.f.    Wald/d.f.  chi pr
Genotype                          1168.72    168         6.96  <0.001

* MESSAGE: chi-square distribution for Wald tests is an asymptotic approximation
(i.e. for large samples) and underestimates the probabilities in other cases.
```

The warning message is discussed further below. With only one fixed model term, the two summary tables contain the same values, and indicate very strong evidence for differences between genotypes (p-value<0.001).

For a single effect, the Wald statistic is equivalent to the square of the t-statistic, calculated by dividing the estimated effect by its standard error. For a set of effects corresponding to a model term, a Wald statistic can be thought of as the sum of squares of the effects weighted by their variance-covariance matrix. The asymptotic reference distribution for the Wald statistic is a chi-square distribution with df equal to the df of the term. This distribution ignores variation associated with estimation of the variance components. Hence, it is analogous to using a Normal rather than a t distribution for a

two-sample test when the variance is unknown. Some caution is therefore required in the use of Wald tests, which tend to be too optimistic, i.e. to give false positive results more often than would be expected (Welham & Thompson, 1997). However, the approximation improves greatly as the residual df associated with the estimated variance-covariance matrix increases. In the context of QTL detection with large numbers of genotypes, the approximation will often be satisfactory.

To avoid this problem, various methods exist to convert, or adjust, Wald statistics into a form that can be comparable to an F distribution, where the denominator df gives a measure of uncertainty in the variance estimation. The most popular method was introduced by Kenward and Roger (1997). This method divides the Wald statistic by the df of the treatment term to give the form of an F statistic. This statistic is compared to an F distribution with numerator df equal to those of the treatment term, and the method provides an estimate for the denominator df. As with the Satterthwaite approximation (*see ANOVA and Design Guide*), these denominator df will often be non-integer. For balanced designs, F tests based on the Kenward-Roger method will be exactly the same as F tests based on the variance ratios in an ANOVA table. By default, Genstat will calculate the approximate F-tests in preference to Wald statistics. This default can be changed using the Options on the Linear Mixed Models menu. Unfortunately, calculation of this adjustment can be costly in terms of memory and occasionally fails, as in the example above. In this case, the Wald tests will be used instead.

The table of marginal Wald or approximate F-tests can be used to implement a process of backwards selection for the fixed model, refitting the model each time a term is dropped. Predictions from the final model can be made using the Predict button, and the fitted means can be graphed using Further Output | Means Plot

10.7 Model checking and goodness of fit

There are several different definitions of residuals from the linear mixed model. Predictors for the model deviations are known as the conditional residuals. Alternatively, marginal residuals can be calculated as the sum of all of the random effects for each observation. Similarly, fitted values for each observation can be calculated using all except the final (deviations) term (which we denote conditional fitted values) or omitting all of the random terms (which we denote marginal fitted values). Both types of residual and fitted value can be obtained from Linear Mixed Models | Save by choosing the terms used to form the residuals - these terms are then excluded from the fitted values (Figure

10.4). Conditional residuals and fitted values are not formed when the model contains negative variance components, as these quantities are undefined in this case.

Figure 10.4: Saving residuals and fitted values. Conditional residuals and fitted values are saved by setting Method for Residuals to Final random term only. Marginal residuals and fitted values are saved by setting Method for Residuals to Combine all random terms.

Assumptions of normality, independence and equal variance can be investigated using these quantities in addition to the BLUPs from the individual random terms (described in Section 10.5).

The residuals and BLUPs are not standardized or pre-whitened. For conditional residuals or BLUPs, we expect that they should reflect the variance model for the random term from which they arise. For a set of independent random effects, it is often useful to

form a histogram or Normal quantile plots of the residuals or BLUPs to assess the approximation to a Normal distribution. For random effects with serial correlation and a common variance, it may be helpful to plot a one- or two-dimensional variogram to assess whether the observed pattern of correlation matches that expected for the model (see Section 3.6).

Construction of the fitted values plot, which plots the residuals against the fitted values to detect variance heterogeneity, requires some thought. The inclusion of random effects in the conditional fitted values means that shrinkage may induce correlation between the conditional residuals and fitted values, which occasionally results in trend in a fitted values plot. If this does occur, it can be investigated by plotting the conditional residuals against the marginal fitted values. If the trend vanishes, then it is an artefact of shrinkage, and can be ignored; if the trend does not vanish, then it indicates some underlying problem with the model. Following an analysis, diagnostics plots for conditional residuals can be generated from the Linear Mixed Models menu using the Further output button and then selecting the Residual Plots button on the Linear Mixed Model Further Output menu (Figure 10.5). The resulting composite set of residual plots for model 1 for the `HEAT05` trial is shown in Figure 10.6. There is no evidence of variance heterogeneity or a non-Normal distribution of the deviations in these plots.
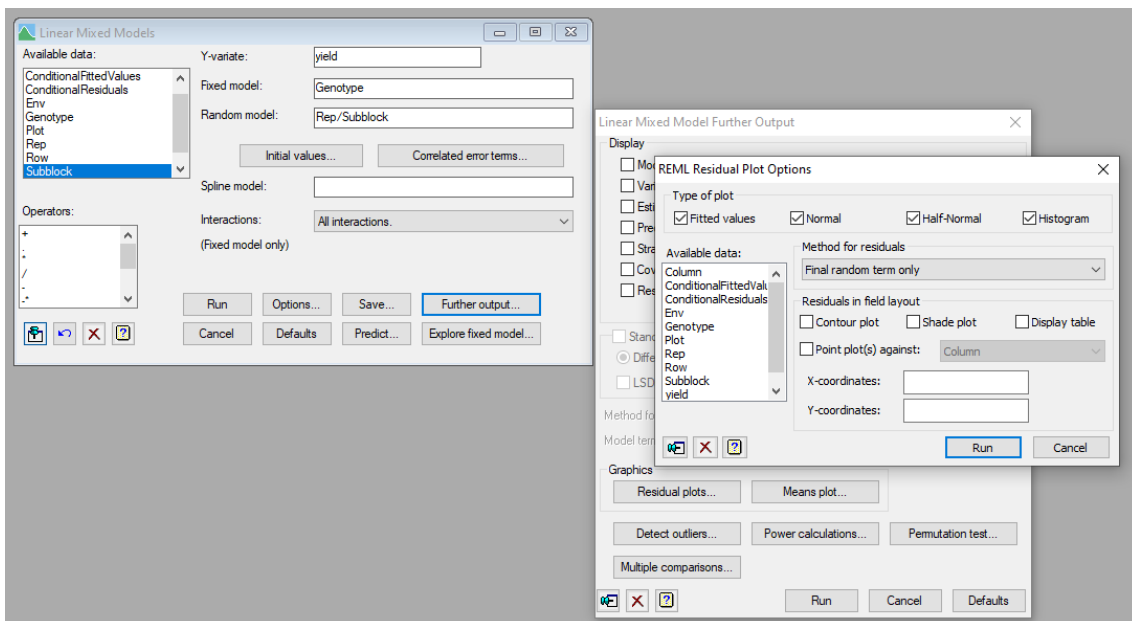

Figure 10.5: Generating conditional residual plots using the Residual Plots button on the Linear Mixed Model Further Output menu.
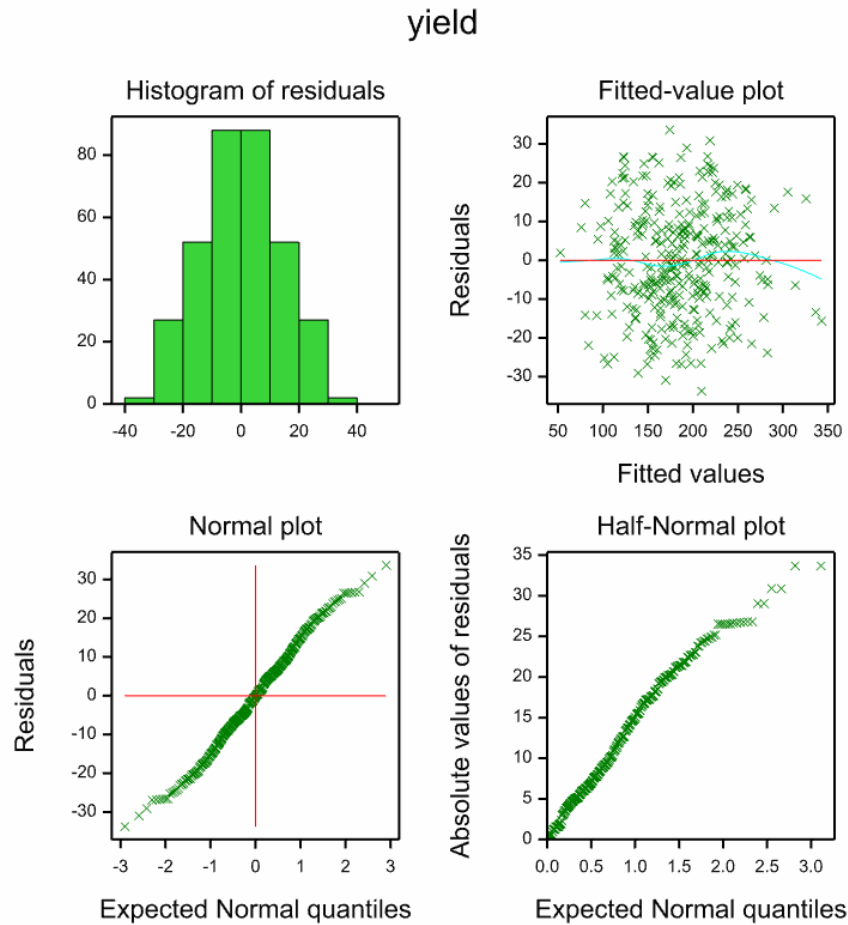
Figure 10.6: Composite set of plots for conditional residuals from the `HEAT05` trial.

There is currently no generally accepted analogue of the adjusted R^2 statistic to quantify the explanatory performance of linear mixed models. It is generally acceptable to state whether fixed terms show evidence of group differences (factors) or linear trend (variates), based on the outcome of Wald or approximate F-tests. However, one approach to calculating the percentage variance accounted for by the fixed model is based on the variance of an observation, as defined in Section 10.2, in terms of the sum of the variance components. The baseline total variance can be calculated as the sum of the variance components when the constant term alone is included in the fixed model. This is compared to the sum of the variance components for the fixed model under consideration, and the percentage reduction can then be considered as the percentage variance accounted for by the fixed model. This statistic can be used to quantify the performance of different fixed models, for a given response and random model.

As an illustration, we calculate the % variance accounted for by the genotype factor in Model 1 for the `HEAT05` trial. If we fit the model

Fixed model:
Random model: `Rep/Subblock`

and ask for a model summary and estimated variance components to be displayed, then we obtain the following output:

```
REML variance components analysis
=================================


Response variate:  yield
Fixed model:       Constant
Random model:      Rep + Rep.Subblock
Number of units:   338

Residual term has been added to model

Sparse algorithm with AI optimisation


Estimated variance components
-----------------------------

Random term              component       s.e.
Rep                         1208.       1743.
Rep.Subblock                 183.         93.


Residual variance model
-----------------------

Term               Model(order)  Parameter      Estimate    s.e.
Residual           Identity      Sigma2            1787.    143.
```

This is a null model, with only the overall constant term in the fixed model. The total variance is equal to the sum of the variance components and the residual variance, and is equal to 1208 + 183 + 1787 = 3178. In Section 10.2, we found that the total variance for Model 1, which included the `Genotype` factor as a fixed term, was 1824. The percentage variance accounted for by adding the genotype term into the model is therefore $100 \times$ (3178-1824) / 3178 = 42.6%.

10.8   A recipe for analysis of linear mixed models

Various different strategies are available for establishing a suitable mixed model for a given data set. The following strategy should lead to a sensible model for a designed experiment:

Step 1: Establish a baseline set of fixed and random terms. The randomization structure of the experimental design should be encompassed in the random terms. The experimental treatments may be included as either fixed or random terms, according to the aims of analysis (see Section 10.1).

Step 2: Run the baseline model, and assess the residuals for evidence of departures from the model assumptions. Variance heterogeneity related to fitted values may be dealt with by use of a suitable transformation of the response variate. Departures such as variance heterogeneity related to treatment groups, or serial correlation, require extension of the random model (Section 10.3). Other common departures include nonlinear trends for explanatory variates, or relationships with factors or variates not included in the model; these may be dealt with by adding suitable terms into the fixed or random models.

Step 3: Extend and refine the random model using the full fixed model - but retain all terms associated with the randomization structure. Assess differences between random models using the deviance or information criteria such as AIC or SIC (Section 10.4). Add fixed terms to account for global trends, e.g. linear trend across a field trial.

Step 4: Once a suitable random model has been determined, refine the fixed model. The table of marginal Wald or approximate F-tests (Section 10.6) can be used to implement a process of backwards selection for the fixed model, refitting the model each time a term is dropped. Make predictions from the final model.

10.9   References

Crainiceanu, C.M., & Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**, 165-185.

Kenward, M.G., & Roger, J.H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**, 983-997.

Patterson, H.D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545-554.

Smith, A.B., Cullis, B.R., & Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *Journal of Agricultural Science*, **143**, 449-462.

Stram, D.O., & Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171-1177.

Welham, S.J., & Thompson, R. (1997). Likelihood ratio tests for fixed model terms using residual maximum likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**, 701-714.

11 Genstat commands

The Genstat QTL system comprises a set of menus and commands to facilitate QTL analysis, bringing together a wide range of statistical techniques. In this Guide, QTL detection has been demonstrated using the menu system. However, all analyses can be achieved using commands in the Genstat language. This chapter introduces the suite of procedures available in Genstat's QTL system, several of which make use of the REML facilities. Full documentation is provided in *Genstat Reference Manual (Part 3 for Procedures)* available via Help | Reference Manual | Procedures (Figure 11.1), and within the Genstat Help System (⬚).
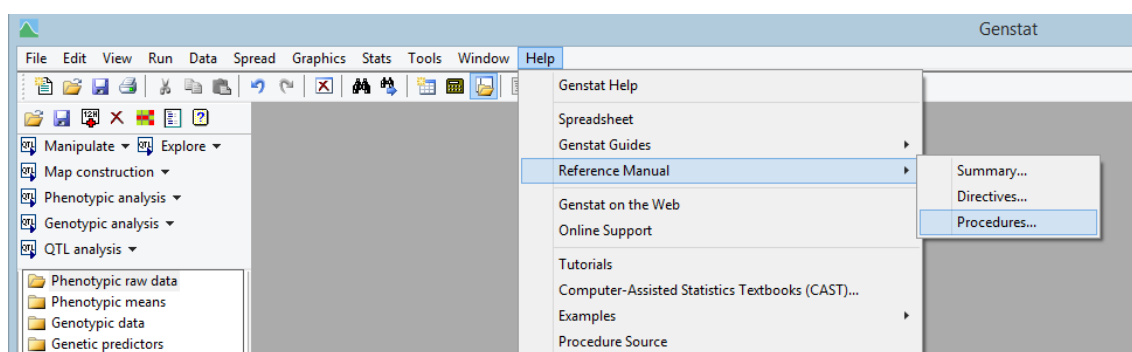

Figure 11.1: Accessing the *Genstat Reference Manual (Part 3 for Procedures)*.

Many of the procedures listed below offer additional options, not specifiable via the QTL menus, which allow for more flexible and/or complex analyses. To learn more about unlocking the full power of the Genstat command language, refer to *An Introduction to the Genstat Command Language*.

Note, use of the menus will generate commands in the Input Log window that can be used to repeat the analysis at a later date, or edited to modify the analysis if desired.

Data import/export

QEXPORT	exports genotypic data for QTL analysis
QFLAPJACK	creates a Flapjack project file from genotypic and phenotypic data
QIMPORT	imports genotypic and phenotypic data for QTL analysis

Data manipulation

QMATCH   matches different data structures to be used in QTL estimation

QMVESTIMATE   replaces missing marker scores using conditional genotypic probabilities

QMVREPLACE   replaces missing marker scores with the mode scores of the most similar genotypes

Data exploration

DQMAP   displays a genetic map

DQMKSCORES   plots a grid of marker scores for genotypes and indicates missing data

QMKDIAGNOSTICS   generates descriptive statistics and diagnostic plots of marker data

QMKRECODE   recodes marker scores into separate alleles

Map Construction

DQRECOMBINATIONS   plots a matrix of recombination frequencies between markers

QLINKAGEGROUPS   forms linkage groups using marker data from experimental populations

QMAP   constructs genetic linkage maps using marker data from experimental populations

QMKSELECT   obtains a representative selection of markers by means of genetic distance sampling or genetic distance optimization

QRECOMBINATIONS   calculates the expected number of recombinations and the recombination frequencies between markers

Phenotypic analysis

VGESELECT   selects the best variance-covariance model for a set of environments

Genotypic analysis

GPREDICTION   produces genomic predictions (breeding values) using phenotypic and marker information

QEIGENANALYSIS   uses principal components analysis and the Tracy-Widom statistic to find the number of significant principal components to represent a set of variables

QGSELECT   obtains a representative selection of genotypes by means of genetic distance sampling or genetic distance optimization

QIBDPROBABILITIES   reads molecular marker data and calculates IBD probabilities

`QKINSHIPMATRIX` forms a kinship matrix from molecular markers

`QLDDECAY` estimates linkage disequilibrium (LD) decay along a chromosome

QTL analysis

`DQMQTLSCAN` plots the results of a genome-wide scan for QTL effects in multi-environment trials

`DQSQTLSCAN` plots the results of a genome-wide scan for QTL effects in single-environment trials

`QBESTGENOTYPES` sorts individuals of a segregating population by their genetic similarity with a target genotype, using the identity by descent (IBD) information at QTL positions

`QCANDIDATES` selects QTLs on the basis of a test statistic profile along the genome

`QDESCRIBE` prints summary statistics of genotypes

`QMASSOCIATION` performs multi-environment marker-trait association analysis in a genetically diverse population using bi-allelic and multi-allelic markers

`QMBACKSELECT` performs a QTL backward selection for loci in multi-environment trials or multiple populations

`QMESTIMATE` calculates QTL effects in multi-environment trials or multiple populations

`QMQTLSCAN` performs a genome-wide scan for QTL effects (Simple and Composite Mapping) in multi-environment trials or multiple populations

`QMTBACKSELECT` performs a QTL backward selection for loci in multi-trait trials

`QMTESTIMATE` calculates QTL effects in multi-trait trials

`QMTQTLSCAN` performs a genome-wide scan for QTL effects (Simple and Composite Interval Mapping) in multi-trait trials

`QMVAF` calculates percentage variance accounted for by QTL effects in a multi-environment analysis

`QREPORT` creates an HTML report from QTL linkage or association analysis results

`QSASSOCIATION` performs marker-trait association analysis in a genetically diverse population using bi-allelic and multi-allelic markers

`QSBACKSELECT` performs a backward selection for loci in single-environment trials

`QSELECTIONINDEX` calculates (molecular) selection indexes by using phenotypic information and/or molecular scores of multiple traits

`QSESTIMATE` calculates QTL effects in single-environment trials

QSIMULATE   simulates marker data and QTL effects for single and multiple environment trials

QSQTLSCAN   performs a genome-wide scan for QTL effects (Simple and Composite Mapping) in single-environment trials

QTHRESHOLD   calculates a threshold to identify a significant QTL