The potential for association mapping from historical trait data in wheat and barley.

Jonathan White.

Imperial College London,
Department of Epidemiology, Public Health and Primary Care

Thesis submitted for examination for the award of the degree of Doctor of Philosophy.

Declaration of Originality

To the best of my knowledge and intention all work presented which is not my own has been appropriately referenced or otherwise acknowledged.

Acknowledgements

To attempt to acknowledge everyone who has helped me during the course of this research would be to risk unintended omission. So apart from colleagues at NIAB and fellow students I would like to specifically acknowledge just three people; my supervisors, Ian Mackay and David Balding, who's guidance was given seemingly without limit of time or patience and my wife, who I simply do not deserve.

The financial support of the following is gratefully acknowledged.

The NIAB Trust.
Crop Evaluation Limited.
The Home Grown Cereals Authority R&D Committee.

Barley genotype data was provided by:

The AGOUEB consortium.

Abstract

Plant breeding continuously produces new cultivars (varieties) of agricultural crops, which are independently evaluated in field trials prior to commercialisation. Variety evaluation has been ongoing in England and Wales since the 1930s and has been a statutory requirement in the UK since 1963. The legacies are seed banks of historic cultivars and corresponding phenotype databases.

The hypothesis to be tested is that this resource, with some genotyping of seed, can be exploited in association mapping experiments. The chosen species represent those with the most complete historic phenotype databases and the least within-variety genetic variation.

The thesis examines the population structure and cryptic relatedness in the available panels with reference to their confounding effects on association mapping. Different analysis methods are considered and mixed effects modelling is identified as the most reliable. Association analysis of Mendelian and quantitative traits is conducted and the results compared with the success rate expected from simulation. Some of the potential QTL identified are shown to be co-incident on genetic maps with QTL identified by others in bi-parental mapping studies, the observed level of co-incidence is found to be significantly greater than would be expected by chance alone.

In conclusion the potential for association mapping of quantitative traits using this experimental model is constrained by the number of varieties for which sufficient phenotype data exists. In wheat the number of available mapped molecular markers is also limiting. Simply inherited traits, however, appear amenable to study in this way.

Larger genotyping arrays are becoming available but simulation shows that larger panels (2x for barley, 5x for wheat) of individuals will be needed if the power of the experiments is to be sufficient to discover, for example, 80% of loci contributing $\geq$10% of variation. The potential to achieve this panel size using advanced inter-crossed populations is briefly discussed.

Table of Contents

1 GLOSSARY....8

2 INTRODUCTION....9

2.1 Hypothesis....9

2.2 Cereal: evolutionary background....9

2.3 The UK cereal crop and genetic resources in agriculture....10

2.4 Plant varieties....13

2.5 Genetics of Wheat and Barley....16

2.6 Wheat and barley varieties in genetic studies....18
2.6.1 Genetic signatures of plant breeding. 19

2.7 Conclusion....25

3 ASSOCIATION MAPPING; THEORY AND PRACTICE. 26

3.1 Linkage or association? 26

3.2 Problems of population structure. 27
3.2.1 Modelling population structure. 28

3.3 Problems of kinship....34
3.3.1 Kinship based on pedigree....35
3.3.2 Kinship from marker genotype data. 36
3.3.3 Boundaries between kinship and population structure. 43
3.3.4 Experimental design for the control of the effects of population structure and kinship. 48
3.3.5 Statistical control of the effects of population structure and kinship.....48

3.4 Performing a test for association. 58

3.5 Statistical Significance and multiple testing....59
3.5.1 Using false discovery rates to determine significance.....59

3.6 Statistical Power. 63

4 ASSOCIATION MAPPING IN PLANT SPECIES: LITERATURE REVIEW. 67
4.1.1 Barley (self pollinating diploid). 67
4.1.2 Arabidopsis (self pollinating diploid)....73
4.1.3 Maize (naturally out pollinating diploid; often inbred by man). 76
4.1.4 Durum Wheat. Self-pollinating allotetraploid. 79
4.1.5 Potato. Out-pollinating autotetraploid. 82
4.1.6 Wheat. Self-pollinating allohexaploid. 85
4.1.7 Plant association genetics conclusion.....90

5 PANELS USED FOR ASSOCIATION MAPPING. 92

5.1 Barley Panel.....92

5.1.1 Germplasm and genotyping. 92
5.1.2 Genetic Map 93
5.1.3 Phenotype. 97
5.1.4 Population substructure. 97

5.2 Wheat Panel for Association Mapping. 101
5.2.1 Phenotype data. 101
5.2.2 Sources of Seed. 106
5.2.3 DArT Genotyping. 107
5.2.4 Genetic Map 123
5.2.5 Population Sub-structure. 137

5.3 Simulated Cereal Population. 141
5.3.1 Simulated Population Development. 141
5.3.2 Comparing the simulated cereal with the wheat and barley populations. 154

6 STATISTICAL POWER IN ASSOCIATION MAPPING. 157

6.1 Estimation of power. 157

6.2 Comparison between experiments. 158
6.2.1 Efficiency score. 159
6.2.2 The power ratio. 160

6.3 Simulated genotypes and phenotypes. 162
6.3.1 Simulated phenotypes 162
6.3.2 Minor Allele Frequency (MAF) in Simulations. 165

6.4 Results 166
6.4.1 Perturbation of existing phenotype. 166
6.4.2 Simulation of phenotypes in wheat, barley and the simulated cereal. 179

6.5 Conclusion 189

7 ASSOCIATION MAPPING: WHEAT PANEL. 191

7.1 Structure and kinship 191
7.1.1 Defining the Q matrix. 195
7.1.2 Defining structure in terms of principal coordinates. 199
7.1.3 Modelling Kinship 202

7.2 Association Mapping. 203
7.2.1 Genomic Control 203

7.3 Structured Association (SA) 210
7.3.1 Significance of association. 220

7.4 Correction using principal coordinates 224
7.4.1 QTL discovery using the PC method. 228

7.5 Mixed effects modelling. 234
7.5.1 Significance of association. 244

7.6 Significance of discoveries 250

7.7 Discussion. 253

8 ASSOCIATION MAPPING: BARLEY PANEL. 256

8.1 Marker and variety panels 256

8.2 Statistical Power. 256

8.3 Phenotypes 257

8.4 Association Mapping. 259
8.4.1 Effect of structure. 259
8.4.2 Modelling structure. 260

8.5 Association Mapping. 265
8.5.1 Genomic Control. 265
8.5.2 Significantly associated markers. 268

8.6 Structured Association. 274
8.6.1 Significantly associated markers. 279

8.7 Correction using principal coordinates. 283
8.7.1 Significantly associated markers. 286

8.8 Correction using mixed effects modelling. 290
8.8.1 Summary of Significant Associations 295

8.9 Confirmation of association mapping discovery and identification of a functional polymorphism for ANT2. 301

8.10 Conclusion. 306

9 DISCUSSION 308

9.1 The experiments. 308

9.2 Comparison with human studies. 309

9.3 Methods, power and false discovery. 312

9.4 Quantitative traits. 315

9.5 Feasibility of association mapping in wheat and barley. 318

9.6 Experimental design to increase power in wheat and barley association mapping. 319

9.7 Beyond association mapping. 321

10 CONCLUSION 323

11 REFERENCES 326

12 APPENDICES. 339

12.1 Appendix 1. The mixed effects model (MM): a summary. 339

12.2 Appendix 2. Performance of regression models in the simulated cereal population with varying numbers of individuals, markers and loci. 349

12.3 Appendix 3. Barley varieties in the AGOUEB Panel. 354

12.4 Appendix 4. List of wheat varieties included in the final association panel. The reference number is that given by the testing authority when the breeder made an application for registration. 358

12.5 Appendix 5. Protocol For Modified Tanskley's DNA Microprep 359

12.6 Appendix 6. Histograms of barley phenotype scores. 365

12.7 Appendix 7. Association plots for barley phenotypes. Association estimated without correction for structure. 371

12.8 Appendix 8. Association plots for barley phenotypes. Association estimated following adjustment for structure using SA + GC. 388

12.9 Appendix 9. Association plots for barley phenotypes. Association estimated with the effect of structure controlled using PCA +GC. 405

12.10 Appendix 10. Association plots for barley phenotypes. Association estimated with the effects of structure corrected using the mixed model. 424

1 Glossary

EST - expressed sequence tag. Refers to a short sequence of DNA used to tag transcribed DNA.

F_{ST}: the fixation index. Can be variously estimated but is essentially one minus the ratio of two estimates of the population genetic variance. The numerator estimate is a pooled estimate over proposed structural groups and the denominator is an estimate made over all observations. Thus as groups approach fixation genetic variation within them falls and the value of F_{ST} will rise.

GLM - the generalized linear model is a flexible generalisation of ordinary least squares regression. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

MAF- minor allele frequency; the frequency with which the minor allele for a marker is observed in the panel.

QQ and PP plots. The distribution of p-values under the null is uniform. For any test statistic there exists a distribution of the quantile which corresponds to the null distribution of p-values. QQ plots are a way to compare the observed distribution of a test statistic with the distribution expected under the null. In a QQ plot the ordered values observed in the experiment are plotted over a corresponding number of ordered values under the null. Positive deviation from y=x is indicative of an inflated test statistic suggesting some or all of the observations lie under an alternate hypothesis. A PP plot is constructed in the same way using $-\log_{10}$(p-values).

Standardised genotype. The process of subtracting the mean allele frequency for a marker from all genotypes for that marker. Missing data may then be replaced with a zero. Genotypes are then divided by the genotypic standard deviation for the marker resulting in a genotype standardised to have mean = 0 and variance =1.

2 Introduction

The original goal of this PhD was to investigate the potential to apply association mapping techniques to wheat using historically recorded phenotype data. During the research a data set of barley varieties became available and it was possible to extend the study to both species.

Association mapping exploits statistical association between genetic variation and phenotypic variation to locate QTL. Thus, in simple terms, the requirement for an association study is a panel of individuals with genotype and phenotype data.

As will be discussed, there exist for crop species seed banks where authentic seed of historical crop varieties (variety defined in 2.4, below) may be obtained as well as historical records from field experiments conducted on these varieties in the past. Thus it appeared that if genotypes could be obtained for the varieties, an association study could be conducted without the need for the collection of additional phenotype data.

2.1 Hypothesis

Successful association mapping studies may be conducted in wheat and barley by exploiting historical varieties and corresponding archived phenotypes.

2.2 Cereal: evolutionary background

The monocotyledonous plants (which include all the cereal species) diverged from other higher plants about 90 million years ago (MYA). Although rice may have diverged from other grasses as long as 40 MYA, wheat and barley are thought to have a shared lineage as recently as 10 MYA. Wheat species recognisable to modern agriculturalists arose less than 0.5 MYA and bread wheat (one of the species used herein) probably arose less than 0.01 MYA (see; Figure 2-1)

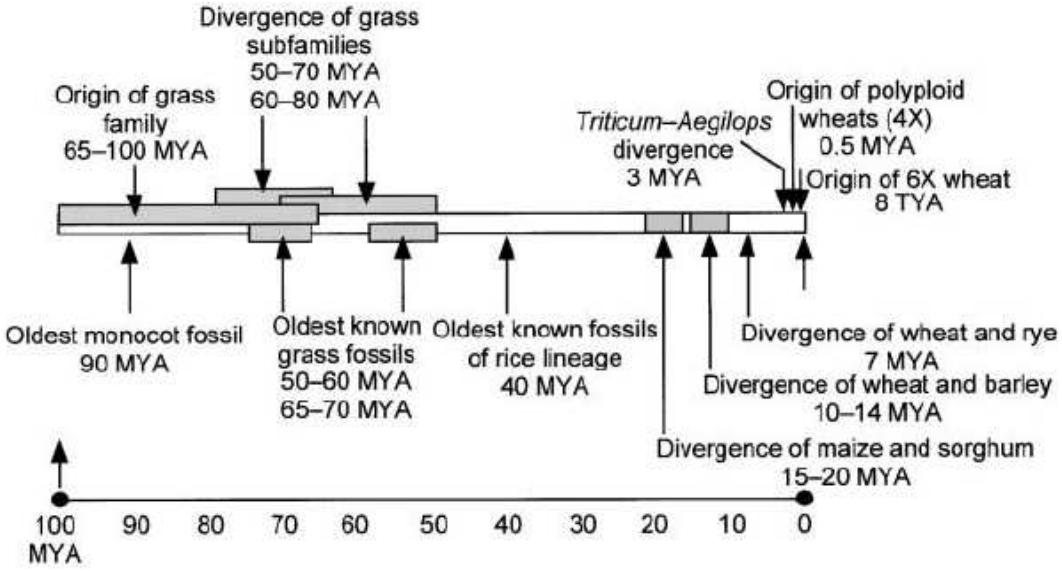


Figure 2-1. The evolutionary history of cereal species. (Reproduced from Gill *et al.*, 2004).

The short period over which the majority of the modern cereal species diversified has the useful consequence that strong synteny persists between them (Devos and Gale, 1997). This may be exploited to overcome some of the limitations which arise because the genomes under study have not yet been sequenced. Where syntenic species with sequenced genomes, such as rice and brachypodium (a group of species closely related to wheat and barley), are available cross-referencing can be used to predict likely genetic structure for the target species.

2.3 The UK cereal crop and genetic resources in agriculture.

The choice of wheat and barley for this research was prompted by their pre-eminent position in UK agricultural research; they are the species for which we have most data. The pre-eminence arises from the economic importance of the species in the UK and the global importance of cereals to human nutrition. To place the crops in context; global cereal production provides mankind with about half its dietary energy (Figure 2-2) and even in first world countries like the United Kingdom direct consumption of cereals still represents the single largest source of energy (Figure 2-3), secondary consumption though animal products probably means the UK also relies on cereals for about half of its dietary energy. In terms of production the UK cereal crop is dominated by wheat (~70% by weight), barley represents 25% of UK production (Figure 2-4). Globally barley is of little importance but wheat is considered as one of the ‘big three’ cereal crops along with maize, and rice (Figure 2-5).

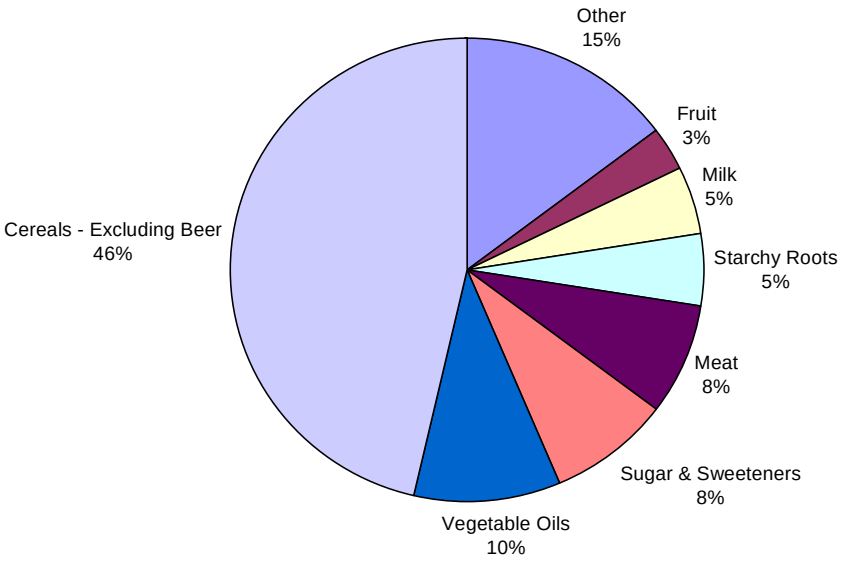


Figure 2-2. Contribution of major food classes to global *per capita* dietary energy consumption (2007). Cereals provide just under 50% of dietary energy. Source: FAO Statistics Division 2010.

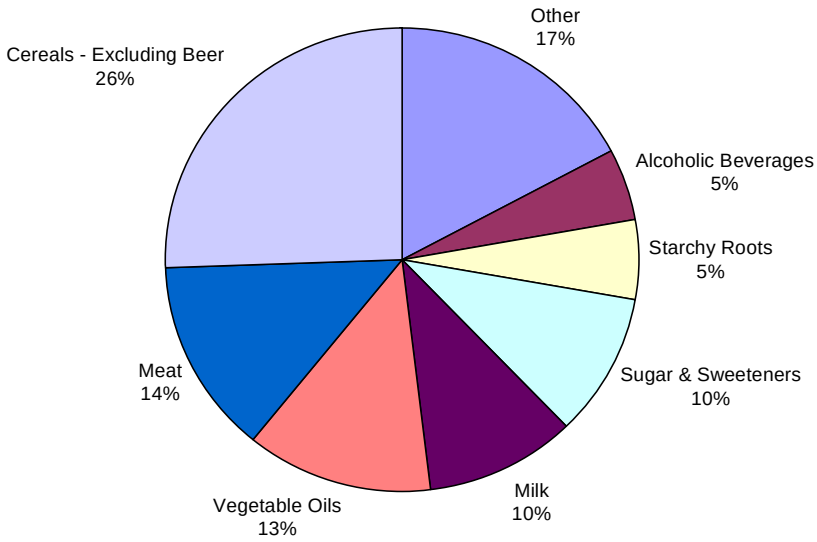


Figure 2-3. Contribution of major food classes to UK *per capita* dietary energy consumption (2007). Direct consumption of cereals provides about 25% of dietary energy. Source: FAO Statistics Division 2010.

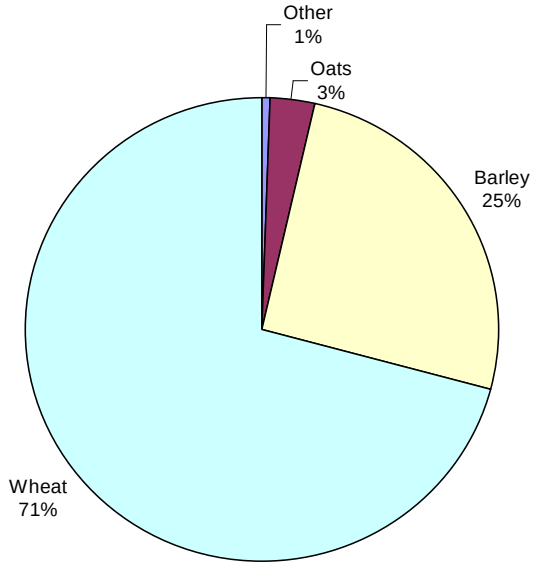


Figure 2-4. Contribution of major cereal species to UK cereal production (2008). Wheat and Barley account for over 90% of production by weight. Source: FAO Statistics Division 2010.

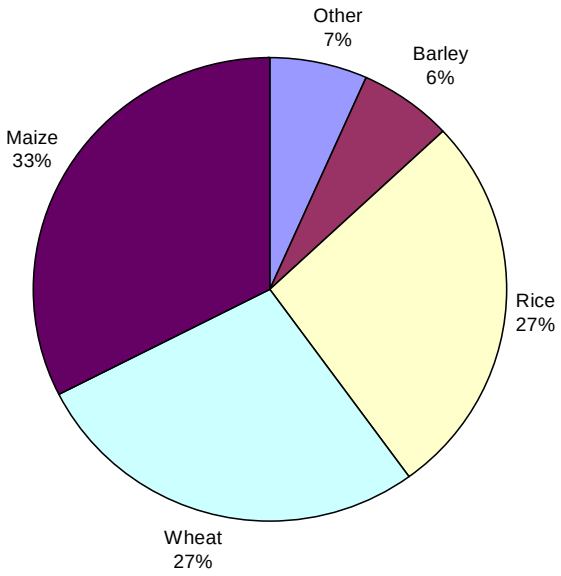


Figure 2-5. Contribution of major cereal species to global cereal production (2008). Wheat, Barley, Rice and Maize together account for over 90% of production by weight. Source: FAO Statistics Division 2010.

Genetically, wheat and barley represent interesting subjects for study; they have both been subject to selection and breeding over the past two centuries and the intensity and sophistication of breeding has accelerated considerably in the past 60 years. The signature of this breeding activity in the genomes will be discussed and it will be found to have consequences for association mapping.

The importance of plant breeding in modern agriculture is hard to over-state. The recent history of wheat production in developed agriculture is often used as a case study exemplifying the 'green revolution' of the 1960s and 1970s. In this period British wheat yields more than doubled and this has been attributed in equal part to improved genetics and improved husbandry (Silvey, 1986). Subsequently the British wheat crop has continued to benefit from plant breeding with continued, more modest, increases in yield and improved quality of flour being the main features (White, 2001; Mackay *et al*., 2011). Barley has also benefited from the attentions of plant breeders although the improvements in yield performance have not been as great as for wheat (Mackay, *et al.*, 2011). This reflects both the more minor status of the barley crop and the fact that much of the breeding effort has been directed towards improving or maintaining its potential as a raw material for the production of alcoholic drinks; some components of the two traits (yield and brewing potential) being negatively correlated with each other.

Crop development by breeding and agronomy is a continuous process since pressure from changes in the grain market and the natural and political environments mean that the nature of the 'ideal' genotype is being constantly re-defined. Any additional knowledge of crop genetics resulting from association mapping which could be used to accelerate the breeding process in pursuit of this changing ideal will have economic benefits to the agricultural industry and improve the food security/quality available to consumers.

2.4 Plant varieties

The International Union for the Protection of New Varieties of Plants (UPOV, Geneva) defines a variety as:

"a plant grouping within a single botanical taxon of the lowest known rank, which grouping, irrespective of whether the conditions for the grant of a breeder's right are fully met, can be

- defined by the expression of the characteristics resulting from a given genotype or combination of genotypes,

- distinguished from any other plant grouping by the expression of at least one of the said characteristics and

- considered as a unit with regard to its suitability for being propagated unchanged;"

However, this definition assumes knowledge of the way in which enhanced genotypes are developed, maintained and exploited.

In modern agriculture farmers tend to grow varieties of crop species which have improved economic performance over the wild type. The improvement in performance has been achieved by selection (pre ~ 1900) and subsequently, breeding following principles rooted in population genetics.

For the majority of UK crops seed is sown (usually autumn or early spring) and the crop is harvested within 12 months. For many crops the harvested material will, biologically speaking, be viable seed, but in an agricultural context this is not regarded as such because its main purpose is as a commodity for sale for food use. However, farmers do frequently save a small proportion of the harvested product and use it as seed for the subsequent season. The process of producing seed on farm does not continue indefinitely; (a) because seed of higher economic potential – new varieties – are introduced to the market annually and in order to remain competitive farmers need to use the best varieties available, and (b) because for out-breeding species segregation means seed does not grow sufficiently true to type for more than one or two generations. Thus seed and variety production are linked together and are a process distinct from crop production.

To supply new varieties and seed there exists a breeding and seed production industry. Plant breeders produce new varieties which potentially offer advantages in terms of (for example) yield, value for end use or disease resistance. The process of producing wheat and barley varieties is one of assembling combinations of useful traits into a single genetic background. Crossing and back-crossing are used to introgress traits or create new trait combinations and selection and further back crossing over about ten generations eventually leads to desirable new combinations of traits in a genetically homogenous population which can be held in a partially fixed state by careful re-selection; ‘the variety’. From this nucleus seed is multiplied

over four or five generations under close quality assurance to produce seed for sale to farmers.

Molecular biology and tissue culture have, in recent years, accelerated this process; for example marker assisted back crossing exploits molecular markers for desirable traits and allows breeders to discard unwanted segregants before the next backcross generation is even sown. This is particularly powerful for recessive traits which can be lost during backcrossing where the trait exists in heterozygous form before being 'recovered' by self fertilisation at a late generation. Genotyping across the whole genome can be used to measure the genetic distance between the segregant and the backcross parent allowing those segregants containing the introgression and most rapidly approaching the desired parental genotype to be selected. Where a marker associated with a trait is used for selection (as opposed to a marker mapping onto a causative polymorphism) additional flanking markers can be used to detect recombination events between the marker and the trait. Tissue culture can be used to grow haploid plants by (for example) embryo rescue. Doubling the ploidy of the haploid plant by colchicine treatment yields universally homozygous individuals which can short-cut the process of producing inbred lines.

The marketing of varieties is controlled by statute (Plant Varieties and Seeds Act, 1964; Plant Varieties Act,1997) which in turn relies on an international treaty (UPOV, 1991) under which the nature of a variety can be defined species by species. The treaty has the main purpose of allowing breeders to register their new varieties as being distinct from all other varieties and so obtain protection for their intellectual property (Plant Breeder's Rights), a secondary benefit is that once a variety has been registered other breeders may use the variety in their own breeding programmes. Uniqueness is demonstrated by recording a suite of botanical characters which are known to be genetically determined and for which the variation due to environment is low. Thus a variety is defined by its morphology with respect to traits which often have no direct relevance to the commercial success of the crop.

To ensure that the registration process is robust new varieties must be uniform for the characters used in distinctness and also be stable for these characters over years. This is routinely tested over two consecutive seasons, the independence from environmental effect means records from a single site may be used for this purpose. If a variety is distinct, uniform and stable it may be sold to farmers but in practice the variety must also demonstrate economic benefit before farmers will buy it. For this reason field trials of candidate varieties are conducted prior to registration and registration is only given to those which show

agronomically useful performance judged relative to their contemporaries. Registration takes place after two years of such trials. However, even at this point there is insufficient information for farmers to make an informed choice between competing offers so independently managed[1] variety evaluation trials are conducted over three further seasons. In the UK the results of both initial and subsequent independent evaluation are used by a committee of experts to select the best varieties available at a given time. This selection is presented as a recommended list of varieties together with a summary of relative performance from all trials. It is probably valid to view the variety testing phase of registration as part of the process of artificial selection following breeding. Certainly the most economically successful varieties contribute more to the subsequent variety pedigrees than do their less successful competitors. Current estimates (Rush, 2010 reporting on an analysis by the British Society of Plant Breeders) for the economic benefit from the introduction of improved varieties of wheat and barley in the UK are about £900m *per annum.*

The process of variety evaluation has been on-going in the UK for nearly a century and it is the historical records from this testing process which provide the phenotypic data for this research (for an overview of the history of the UK variety testing system see; Wellington and Silvey, 1997). Practical issues arising from this are the extent to which data are complete and digitised and the further extent of un-digitised paper records which may, as part of this research, be digitally recorded.

2.5 Genetics of Wheat and Barley

Common (or bread) wheat, *Triticum aestivum (6n),* is an allohexaploid resulting from a natural hybridisation event between *T. turgidum (4n)* and *Aegilops tauschii (2n*). *T. turgidum* itself is the allotetraploid progeny of *T. urartu (2n)* and an unidentified diploid of the genus *Aegilops* thought to be closely related to *Ag Speltoides (2n)*. The allohexaploid nature of wheat is reflected in its chromosome structure where n = 7 for both bread wheat and each of the progenitor species. Thus the wheat genome consists of three parental genomes conventionally denoted A (*T. uratu*), B (*Ag. spp.)* and D (*Ag. tauchii*) (Huang, *et al.* 2004).

[1] Managed by Crop Evaluation Limited, a subsidiary company of the Home Grown Cereals Authority. Trials are funded from a small levy paid by growers on each tonne of grain they sell.

Wheat is predominantly cleistogamous[1] in breeding habit such that out pollination usually represents less than 1% of pollination events, consequently, even in wild accessions, the genotype at almost all loci is homozygous. For plant breeders the inherent absence of inbreeding depression simplifies the production of stable varieties.

The close synteny which exists within the *Poaceae*[2] family (Devos and Gale, 1997) is particularly evident between the three genomes of wheat where many traits are controlled by the interaction between copies of what are generally thought to be the same genes present in some or all of the contributing genomes; e.g. Genes conferring insensitivity to gibberellic acid, Rht-D1 and Rht-B1, which in various combinations of allelic forms, can reduce the height of wheat plants in an almost continuous range down to less than 50% of that of the 'wild type' (Flintham, *et al.*, 1997).

The wheat genome, at 16 gb, is one of the largest in the *Poaceae* family however it seems unlikely that the wheat genomes individually contain substantially more genetic information than, for example, rice. The size of the wheat genome is attributed to the presence of large amounts of intronic and other non-coding DNA (Linkiewicz *et al*., 2004, Gill *et al.* 2004).

Thus the wheat genome represents considerable challenges to geneticists:

- Efforts to sequence the wheat genome are complicated by its large size, the need to deal with each genome separately and the long sections of repetitive DNA which separate genic regions.
- Molecular markers for use in wheat must be genome specific if they are to be useful in marker assisted selection or marker assisted back crossing.
- The plasticity of wheat phenotype (see example of plant height, above) is thought to reflect interactions between the three genomes, which also leads to considerable complexity in the mode of expression of many genotypes.

Barley (*Hordeum vulgare*) is a diploid (n = 7) species and as such represents a simpler subject for genetic study than wheat. However its genome size is large (5.3 gb, Wicker *et al*., 2008) due to structural and repeated DNA. Like wheat, barley is largely self pollinating (estimates by Giles (1989) suggest 5% out-pollination is typical under UK conditions) and genetic monoculture make these rare out-pollination events almost equivalent to self pollination.

[1] Pollination takes place before the flowering structure opens - making self-pollination the norm.
[2] *Poaceae*; family of monocotyledonous plants containing the cereals.

2.6 Wheat and barley varieties in genetic studies

Bringing together the genetic and agronomic aspects of the crops; the species are predominantly self pollinating which results in genotypes homozygous at almost all loci. It is important to remember that even when a plant does out-pollinate the probability is very high that it will cross with plants immediately adjacent in the field. This can be exploited in seed production since if:

- land free of seed spilt from earlier crops is always used;
- seed crops are spatially isolated from other crops of the same species;
- seed crops are inspected by botanists to ensure that the morphology of the plants conforms to the description recorded at registration;
- there are no plants of other varieties evident in the seed crop;

Then the result will be a genetic mono-culture and wheat and barley varieties will breed true over many generations. The implication is that, for most purposes, varieties of both species may be regarded as universally homozygous and genetically homogenous, consequently individuals in a plant genetic study using wheat and barley varieties are not individual organisms but the varieties themselves. In addition the universal homozygosity means genotypes can be conveniently recorded as haplotypes and, indeed, all analyses in this thesis have been conducted on the data expressed in this way.

Thus the number of individuals available for study will be restricted to the number of varieties developed and stored in seed banks, and in the context of this research, further constrained by the number of varieties for which usable historical phenotype data are available. However, the study of plant varieties benefits from at least two advantages over the study of human subjects; the number of organisms within a variety is, for all practical purposes, infinite and there are few ethical constraints on the types of experiment to which a plant may be subjected. This means the same genotype can be tested at multiple locations over multiple seasons, the produce from a single experiment may be divided up and destructively tested in a number of ways and, if the subject of the experiment were genetic susceptibility to disease then plants can be artificially infected and left to die untreated if that provides useful data. So restrictions on numbers of individuals is at least partially compensated by the accuracy with which the genetic and environmental components of

phenotypic variation may be determined and the ability to conduct multiple parallel destructive tests on the same individual.

2.6.1 Genetic signatures of plant breeding.

Studies of association mapping in human populations can make a default assumption that mating has been at random, population sub-structure may subsequently be found and quantified but, as was the case in the The Wellcome Trust Case Control Consortium Study (2007) there may be little need to control for the 'hidden variable' of structure in regression analysis.

In contrast in crop species it is reasonable to assume that plant breeding activity will have left strong signatures in the genome and that these may translate into population structure. Here I outline the potential causes of this substructure in the species under consideration, statistical analysis of structure will be conducted and reported in Chapter 3.

2.6.1.1 Kinship and phenotype may have shared directions in time.

There is a fundamental difference between panels of crop varieties and panels of human subjects; in crop varieties there is a direction to breeding (i.e. artificial selection) over time such that the varieties grown in one decade are likely to be more phenotypically and genotypically similar to each other than when comparisons are made between varieties bred in different decades. Human evolution, in contrast, lacks direction over short time frames (~ decades) and we would be surprised to find that, for example, a sample of ethnically matched British men aged 20 were genetically more similar to each other than they were to a similar sample of British men aged 80. Figure 2-6 shows, in the UK wheat panel described by White *et al.* (2008), that genetic distance between varieties increases with the difference in registration date between varieties. In a QQ plot of standardised residuals there is evidence of an excess of negative values and the variation in residuals falls with increasing age difference (data not presented). This suggests a departure from the assumptions underlying Pearson's product moment correlation coefficient so I report the result using ranked correlation; in practice the results from either test are identical.

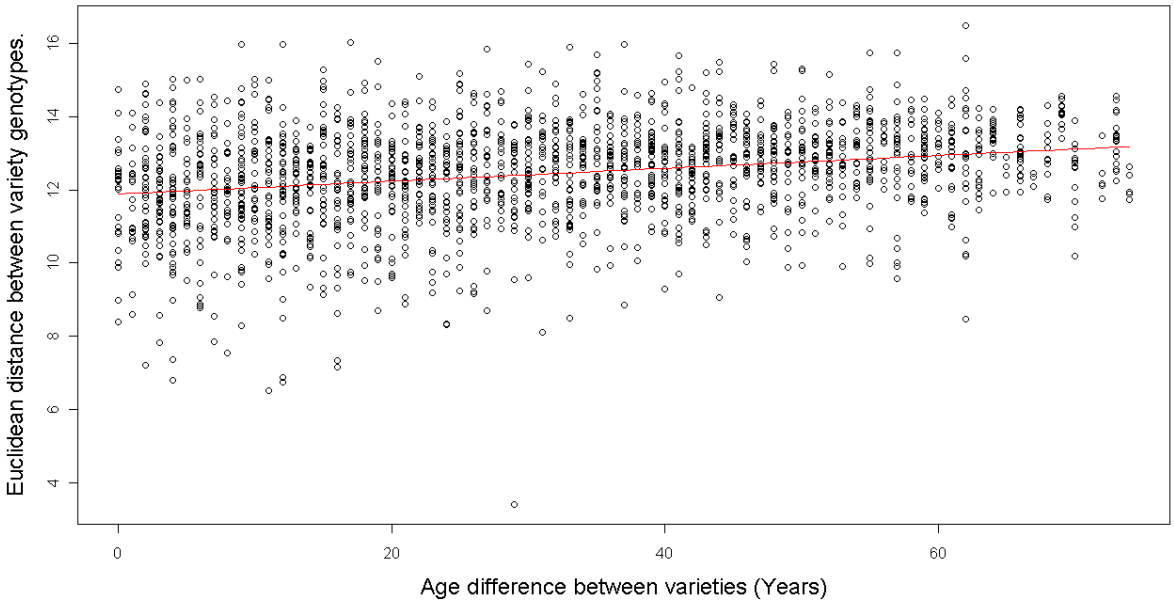


Figure 2-6. The genetic distance between varieties is correlated (rank correlation; $P<10^{-39}$; $R^2=0.09$ (95% confidence interval: 0.069-0.108) with the age difference between the varieties based on year of first registration. Estimated from a panel of 93 UK wheat varieties registered in the period 1931 – 2005 and genotyped with 527 DArT markers. Data have been re-sampled with replacement to give equal numbers of pair-wise comparisons at each age difference interval, (the quoted p-value is the maximum observed in 1000 re-samplings, the confidence interval on R^2 is estimated empirically from the results of re-sampling).

Because there is a weak but highly significant trend in genetic distance over time difference, there is a danger that genetic similarities between contemporaneous varieties will associate with any phenotype which varies over time and potentially result in statistically significant associations, some of which result from markers in close proximity to QTL and others which are simply an artefact of co-ancestry. In this context I note that Mackay *et al*. (2011) report just such a variation in the genetic component of yield for both wheat and barley over this time period.

2.6.1.2 Sub-populations may exist in variety panels.

Genetic drift within populations isolated from each other can account for gross differences between populations without the need to invoke the effects of selection or bottlenecks from founder population sizes (Wright, 1943, 1951). Plant breeders control the breeding of the lines in their breeding pools and so may impose genetic isolation between groups. Where a

crop species has multiple markets plant breeders may choose to create a breeding pool specific to each. When this happens there is a strong likelihood that sub-population structure in the crop will result because the pools may:

- have few founders;
- be selected for differing ideal phenotypes;
- drift apart from each other due to stochastic processes;
- have low migration rates between groups.

In both wheat and barley there is potential for the development of sub-populations because the crops are sold into several markets each with different requirements:

1. Farmers may wish to sow cereal seed in the autumn or the spring; the first reason for this is the date that the preceding crop is harvested; late harvested crops such as sugar beet may be followed by a spring sown crop whereas crops, like oilseed rape or cereals, harvested in August leave land clear for re-sowing in the autumn. Autumn sown crops tend to yield more highly than spring sown crops because they have a longer period of growth. However, the quality of grain (wheat for bread making, barley for brewing) tends to be inversely related to grain yield because well filled grains are higher in starch content and (in the case of wheat) proportionally lower in proteins important for the more specialised markets; well filled barley grains tend to have an excess of storage proteins responsible for haze (lack of clarity) in beer and produce insufficient enzymes during malting to fully convert the starch to sugar, which is commercially inefficient. Grain for specialist markets generally achieves higher prices (15% higher for milling wheat. Source: Home Grown Cereal Authority, July 2010 and 13% for malting barley. Source: Crisp Malting Group, Feb 2010). Thus farmers choose a variety based on practical aspects of farm management, its potential to yield, the potential price and the potential gross margin (price x yield – cost of production).

   Varieties for autumn sowing must tolerate cold and, more importantly, they must delay flowering until after the winter (there is a danger that a crop sown in a mild autumn could initiate flowering before winter, the consequence could be total loss due to destruction of flower buds by frost). Spring sown crops must progress to flowering without delay if they are to be ready to harvest before autumn. The suppression of

flowering until after winter is controlled by *VRN* (vernalisation) genes (Amasino, 2005), *VRN1* being almost universally used in British wheat and in these genetic backgrounds the trait behaves as a dominant Mendelian character. In contrast, frost tolerance is a complex quantitative character (Sutka, 2001. Sutka and Snape, 1989).

The consequence of needing to work with a genetic background which is suitable for winter or spring sowing is that breeders rarely make crosses between the two pools since to do so would incur an additional cost of backcrossing to a 'winter' or 'spring' background. It is reasonable to expect, therefore, that winter and spring cereals will exhibit some genetic differentiation and behave as sub-populations within combined panels.

2. Barley exists in two botanically different forms, termed 6-row and 2-row, which refers to the number of rows of grains in the ear. Until relatively recently taxonomists considered these forms to be separate species (*H. vulgare* and *H. distichim* respectively). In the last century, however, it was determined that a single major gene is responsible for the difference (Engledow, 1920) and the two forms are now classified as *H. vulgare f. hexastichon* and *H. vulgare f. distichon.* On average the 6-row form yields more highly than the 2-row form but the grain produced are of variable size. In N.W. Europe this heterogeneity in grain size is considered detrimental in the production of malt for brewing so 6-row barley is regarded as only suitable for animal feed. However, different brewing technology in N. America allows the use of 6-row barley in malt production so this distinction in quality is not absolute.

   Breeders have invested considerable effort in producing 2-row barley suitable for European style malting; this effort has not been directed at 6-row barley which has been selected predominantly for yield. The consequence is two breeding pools, migration between the pools is rare for similar reasons to those outlined for vernalisation and frost tolerance. Again, it would not be surprising to find that 6-row and 2-row barleys form sub-populations in combined panels.

3. One technique in conventional plant breeding is the introgression of traits from related plant species. In wheat there has been an important introgression from rye chromosome 1 into wheat chromosome 1B (1B1R translocation) conferring resistance to the fungal disease 'yellow rust' (Bonjean and Angus, 2001) and also giving improved drought tolerance (Waines and Ehdaie, 2007). However, wheat containing

this translocation has proved unsuitable for bread making due to the sticky character of the dough and a reduced gluten content (Dahaliwal, *et al*., 1987; Waines and Ehdaie, 2007).

The introgression has replaced much of the short arm of chromosome 1B and since it originates from a single intra-specific cross the genetic diversity of the introgressed fragment is very low. Recombination between the introgression and wheat chromosome 1B has proved to be a very rare event so the introgressed DNA persists virtually unchanged.

The combination of poor milling quality with potentially increased yield has led to the inclusion of 1B1R translocated wheat in breeding pools aiming at the animal feed market but the translocation has been excluded from pools aimed at the milling sector. It seems likely that this segregation may be reflected in population sub-structure.

For illustration, I introduce here two figures (Figure 2-7 and Figure 2-8) which are fully explained, together with more detailed analysis of population substructure, in 5.1.4. These figures show subpopulation structure readily apparent in the first two principal coordinates of genetic correlation matrices of the wheat and barley panels derived from marker data.

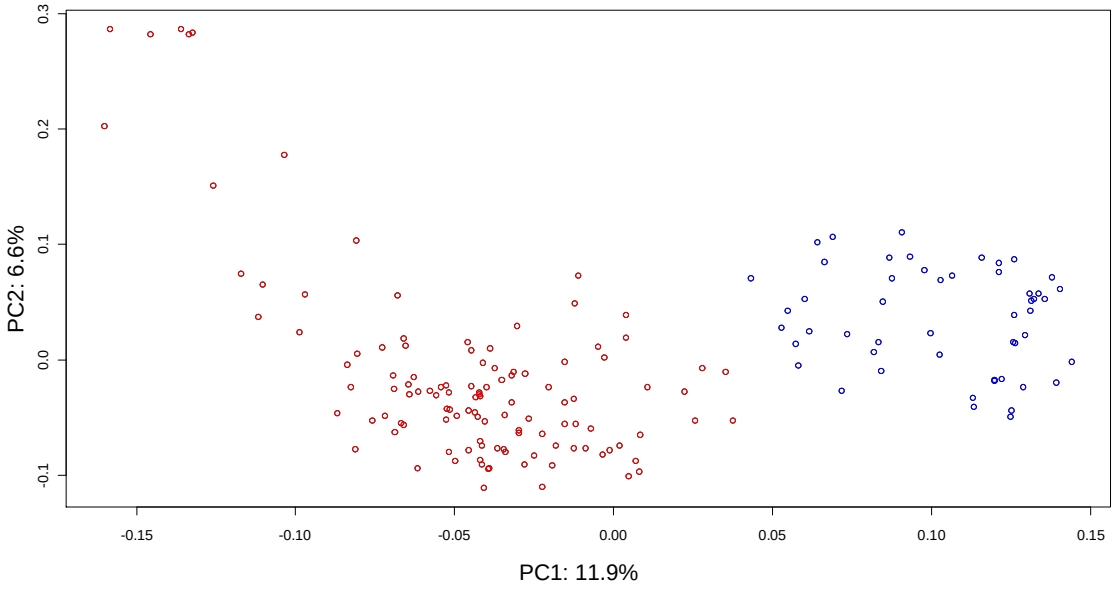


Figure 2-7. Wheat panel: First two principal coordinates of a genetic correlation matrix between individuals. Red circles represent wheat varieties which do not contain the 1B1R translocation, blue circles contain the translocation.

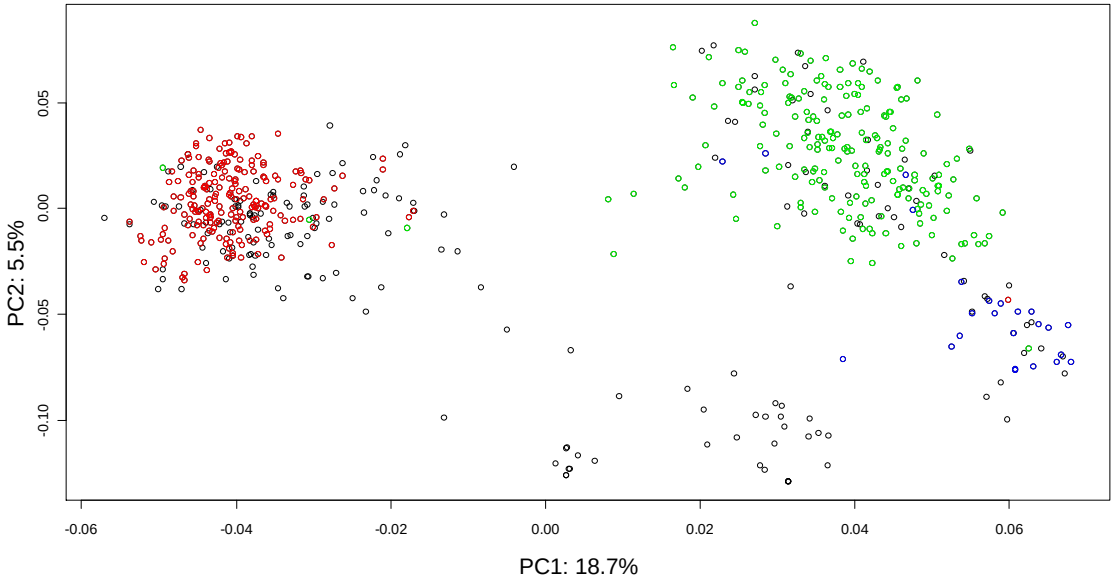


Figure 2-8. Barley panel: First two principal coordinates of a matrix of genetic correlation between individuals. Colours represent; winter/2-row : green; winter/6-row : blue; spring/2-row : red; phenotype not available : black.

2.7 Conclusion

Individual wheat and barley varieties may be regarded as individuals for the purpose of association mapping. The maintenance of varieties by a process of selection to a defined morphology coupled with the in-breeding nature of wheat and barley leads to stability of genotype over generations and genetic homogeneity within a variety.

The source of seed for this research will be important; varieties maintained informally by farmers would be unsuitable since there is no certainty that errors in identification wont have occurred. Indeed, it might be preferable to obtain seed which was placed in storage at the time the variety was current rather than use seed which has been re-generated annually. Thus reference seed-banks, where possible, should be the source of genetic material.

Phenotypic data has been recorded on varieties in evaluation trials annually for nearly a century. Establishing the completeness of digital records and the scope for digitising additional paper records will be an important step in this research.

Since genetic homogeneity is expected it will not be necessary for genotyping to extend to multiple accessions from the same variety, however some replication of some varieties would serve to confirm the assumed homogeneity.

There is good reason to expect population substructure within panels of wheat and barley and there evidence of kinship moving in parallel with phenotype over time. Both these factors will be ‘hidden variables’ in any subsequent regression model.

3 Association Mapping; theory and practice.

In this chapter I will give an explanation of the process of association mapping. I will describe the confounding effects of structure and outline analytical solutions from the literature.

3.1 Linkage or association?

The terms linkage mapping, association mapping and linkage disequilibrium mapping need some definition.

Linkage mapping attempts to locate QTL (Quantitative Trait Loci) by studying the transmission of marker alleles through a pedigree in concert with a particular phenotype. The null hypothesis in this case being random assortment between marker alleles and phenotypic state. Where a marker allele occurs together with a phenotypic state more frequently than expected under this hypothesis there is evidence that the two are in linked, due to proximity of the marker to a relevant gene. The frequency of recombination between two features in the genome is the basis of the calculation of genetic distance; frequent recombination suggests the two features are widely spaced. Because relatively short pedigrees are used (in humans because of long generation times, in other species because of cost) the precision (in terms of genetic distance) is often poor (because relatively few recombination events are observed). However, the accuracy, in terms of correctly locating a QTL to chromosome region, is good (because the experiment is internally controlled against confounding factors).

The terms **association mapping** and **linkage disequilibrium mapping** seem nowadays to be used interchangeably with association mapping becoming more common. Association mapping may be used after linkage mapping to overcome the limited precision of the latter or may be used to discover associations *de novo* in genome-wide association studies (GWAS). With association mapping, populations of individuals are chosen which exhibit the phenotype of interest at a range of levels and genotyping is conducted at high resolution in all individuals. Against a null hypothesis of no association[1] with phenotype, variation in each marker genotype is tested in turn for statistically significant association with trait variation. Association in these circumstances suggests proximity to the gene (if it has already been established by linkage mapping) or QTL and the highest statistical significance of association

[1] The null hypothesis throughout this thesis will be one of no association unless otherwise indicated.

is often used as a guide to markers closest to the relevant feature. Further iterations of testing with additional genotyping in the region of a potential discovery can be used to try to home in on potentially causative SNPs.

It should be noted that the correlation exploited in association mapping is a manifestation of linkage between genotyped marker and causal locus. Where LD is strong in the region of the causal locus the association may be easily detected but with poor resolution. In regions where LD is weak there is the potential for high precision mapping but *de novo* detection of association may be difficult.

Nonetheless, recent advances in molecular biology mean that the high resolution genotyping, formerly directed at sites within the genome where a QTL had been identified by linkage studies, can now be conducted across the entire genome. This gives the opportunity to cut the cost of trait mapping by bypassing the linkage mapping step and considerably reducing the need for pedigreed populations and family based studies. In human genetic epidemiology this approach has been extremely successful with large studies ($\approx 10^5$ individuals and $\approx 10^6$ markers) allowing multiple and verifiable QTL for a particular disease risk to be mapped (for example; Welcome Trust Case Control Consortium, 2007). As indicated (2.6) the numbers of individuals in a plant study based on varieties will be restricted but the genotypic contribution to phenotype may be determined quite accurately. Marker numbers in plant studies are currently lower than those in human studies by a factor of about 1000 (e.g. Cockram, *et al*. 2010 (1111 markers), Cockram, *et al*. 2008 (204 markers), Aranzana *et al*. 2005 (186 markers), Kraakman *et al*. 2004 (236 markers), Yu *et al*. 2006 (553 markers)). In wheat the constraint introduced by the requirement to use genome specific markers also reduces the number of molecular markers available.

3.2 Problems of population structure.

At first sight the technique of association mapping appears quite simple. It relies on the measurement of association between a suite of mapped genetic markers and a phenotypic character using a population of individuals chosen to exhibit the phenotype at a range of levels of expression: where the occurrence of an allele can be shown to correlate with the expression of the trait an association is recorded. Various statistics (see 3.4) can be used to test the significance of observed association in the genotyped sample against a null hypothesis of no association in the population. Statistically significant association should

indicate close genetic proximity and allow the localisation of genes which contribute to the observed phenotype. Also, markers associated with these genes may have value for diagnostic - predictive - purposes.

In practice, however, the genuine associations are often masked by spurious associations which arise from the recent genetic history of the population; its genetic structure. Structure, in this context, relates to genetic differentiation of groups within the population (e.g. controlled breeding with a small founder population; genetic isolation due for example to geography or plant breeding, distinct migration patterns). This leads to allelic frequency differences between sub-populations that may appear to be associated with phenotypic differences if these differences also exist between the sub-populations. Structure is not always evident from ascertainment information. Pritchard and Donnelly (2001) and Astle and Balding (2009) discuss these issues in more detail.

3.2.1 Modelling population structure.

3.2.1.1 Principal coordinates.

In the context of association mapping population structure may be modelled using the first few principal coordinates of a genetic proximity matrix (Zhang *et al*. 2003; Price *et al.* 2006).

Using a correlation matrix calculated from marker genotypes standardised to have mean zero and unit variance we have:

$$\hat{K} = \boldsymbol{X}\boldsymbol{X}^T/L$$

where $\hat{K}$ is a matrix of the pairwise genetic correlation between individuals in the panel, $\boldsymbol{X}$ is the standardised genotype in matrix form (markers in columns) and L is the number of markers in the panel.

$\hat{K}$ may be decomposed into eigenvectors and eigenvalues such that:

$$\hat{K} = v\Lambda v^T$$

where the columns of v are the eigenvectors of $\hat{K}$ and Λ is a diagonal matrix of eignvalues.

Principle Coordinate Analysis and Eigenvectors

Principle coordinate analysis aims to summarise the variation in a set of correlated variables in terms of a new set of uncorrelated variables. In this context the correlation between individuals in an association panel, $\hat{K}$, is complex but could be simplified into vectors (***eigenvectors***), a few of which might contain most of the useful information.

For a matrix of pair-wise correlation between *n* individuals ($\hat{K}$) the total information in the matrix could be summarised in *n* eigenvectors (k_1 to k_n) such that:

$$\hat{K} = \sum_{i=1}^{n} k_i k_i^T$$

The first vector, k_1, is found such that the sum of the squared residuals between $\hat{K}$ and $k_1 k_1^T$ are a minimum. The second vector is calculated in the same way but on a matrix of residuals. Thus each vector in turn accounts for a portion of the original variance and over all vectors we have $\hat{K} = \sum_{i=1}^{n} k_i k_i^T$.

One constraint is placed on the choice of vector; each eigenvector must be orthogonal[1] to all others. Conventionally vectors are also re-scaled such that the sum of squares of the vector is unity (**a**'**a** =1). The scaling factor used is simply the square root of the sum of squares and is the ***eigenvalue*** corresponding to the eigenvector.

The method for finding the vectors subject to these conditions is one of optimisation which, for simple illustrative examples, can be demonstrated with (for example) the MS excel '*solver*' function or the R optimiser '*optim*'. In practice the R command '>eigen(X)' returns the full decomposition of **X** and this has been the method used throughout this thesis.

---

[1] Vectors of length *n* describe direction and length of lines in *n* dimensional space. Orthogonal vectors describe lines which lie at right angles and have the property that **a**'**b** = 0. For example; if $\mathrm{a} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\mathrm{b} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ then it is clear that the vectors are orthogonal in 2d space and **a**'**b** =0.

Since the confounding factor 'structure' in association can be thought of as a form of clustering of similar individuals the first few principal coordinates will often summarise these major features of the data (see; 5.1.4 and 5.2.5 for examples) and give a convenient mathematical summary of how the individuals are related to each other within and between clusters. This mathematical summary can be incorporated as a covariate in regression analysis (3.3.5.3) to correct for the spurious effects of structure on marker-trait associations.

3.2.1.2 Island model for structure.

A conceptual framework for the origins of structure is given by the island model; the population is considered to contain sub-populations with limited migration between them, perhaps related to distance, so that allele frequencies can differ markedly between sub-populations.

Within a sub-population there will be a natural tendency for Hardy Weinberg equilibrium to increase and linkage disequilibrium to decrease over time. Factors such as selection and mutation also affect allele frequencies.

The island model gives not only a conceptual framework but also a basis for inferring structure in a dataset for which there may be no prior knowledge of genetic history or sample origin. In simple terms, individuals are assigned to notional sub-populations in such a way as to minimise both LD and HWD.

The most commonly used software for making these assignments today is STRUCTURE (Pritchard *et al*. 2000; Faluch *et al.* 2003, Faluch *et al.* 2007) (download from: http://pritch.bsd.uchicago.edu/software.html). Essentially this software, given a number of sub-populations (*k*) specified by the user, works within a Bayesian framework to find a maximum likelihood solution in which individuals are assigned to subpopulations in such a way as to optimise the equilibrium constraints already stated.

When first published the software assigned individuals unambiguously to single hypothetical historical populations. Subsequently the method has been made more sophisticated by including admixture between sub-populations. The output from STRUCTURE is a matrix (**Q**) of sub-population assignment for each individual; this is either an unambiguous assignment (as in the first iteration of the method) or a matrix of fractional sub-population membership.

The mathematical representation of structure given by **Q** can be incorporated in regression models to correct for spurious association in association mapping (3.3.5.2).

The principle uncertainty in the use of STRUCTURE is the decision as to the value for *k*. To a large extent the decision rests on how **Q** is to be used. For example, suppose **Q** is to be a covariate in regression correcting for the effects of population structure in association analysis for a given trait. A pragmatic solution is simply to model phenotypic variation using **Q** estimated with increasing values of *k;* when the addition of further sub-populations fails to

reduce the residual error there is no reason to increase *k* further. It should be noted, however, that this approach gives a solution which is correct for the phenotype in question but the result conveys no general information about the historic sub-population structure, only how it correlates with the specific phenotype. I separately report the use of this approach in Cockram *et al.* (2008).

A more general value of *k* may be obtained by considering features of the sub-structure model. In the manual to STRUCTURE the authors suggest that when the likelihood of the data ln(P(D))[1] is maximised the optimal value of *k* has been found. However Kraakman, *et al.* 2004 reports that in a panel of barley varieties the estimate of likelihood continues to rise indefinitely. I note similar effects in (7.1.1) while data below (Figure 3-8) suggest this may be due to STRUCTURE finding signatures from recent kinship when the major population features have been accounted for. Evanno *et al.* (2005) suggest the second derivative of the change in log likelihood divided by the standard deviation of the log likelihood at each value of *k* (referred to as ΔK) is maximal at the 'true' value for *k* and the authors demonstrate this in simulated data using an island model with 5 sub-populations. Recently associates and I (Campana *et al*,. 2010) suggested that F_{ST} might be used as a metric for deciding on a general value for *k* since one would expect F_{ST} to rise as the model of structure improves. We propose the ΔF_{ST} statistic, which is calculated analogously to ΔK (Evanno et al. 2005), as a guide and report that it gives similar results:

$$\Delta F_{ST} = \frac{|\bar{F}_{K+1} - 2\bar{F}_K + \bar{F}_{K-1}|}{s(F_K)}$$

$\overline{F}_K$ is the mean F_{ST} calculated from the F_{ST} values calculated for *k* sub-populations (F_K). The second-order rate-of-change of $\overline{F}_K$ is then divided by the sample standard deviation (s) of F_K to normalize the dispersion of the data around successive values of $\overline{F}_K$. Using F_{ST} in this way has the advantage that it makes the decision on the value of *k* using an index of fixation – which is the underlying mechanism in play – rather than the plausibility of a model which has a range of underlying assumptions, none of which can be entirely true.

These general solutions to the question of number of sub-populations are perhaps more acutely relevant to questions of population history; molecular or genetic archaeology where

[1] Ln(P(D)) is the log likelihood of the model using the available genetic data and assuming *k* sub-populations

the technique has acquired a significant following. I observe, however, that finding parameters by maximum likelihoods is a self referential process, indeed it has some characteristics of a self-fulfilling prophecy in that there will always be a most likely solution. That solution, of itself, is not evidence for a particular population history it would be safest used as evidence to support other hypotheses or to offer guidance as to choice between competing theories.

An important question when modelling with STRUCTURE is the number of cycles of burnin (a pre-analysis phase in which the software determined the Baysian priors for the simulation) and MCMC allowed for the model to stabilise. The default values within the software are 10,000 of each but this is inadequate for all but the smallest datasets since replicates under these conditions give neither reproducible values of ln(P(D)) nor reproducible allocations of individuals to sub-populations, in Cockram *et al.* (2008) I allude to the problem. Since it is important to have a final Q matrix to use we chose (in that paper) to compare replicate runs of structure at a given value of *k* by correlating all sub-population allocations in one replicate with all allocations in the other, recording the maximum correlation obtained for each of the *k* sub-populations. Where the average of these *k* maximum correlations exceeds a given value (I used 0.999) the runs are considered duplicates and this is an indication of a stable solution. In Campana *et al.* 2010 we provide software for this type of comparison. **Figure 3-1** shows the scatter of maximum correlations for minimal cycles of model fitting and for a typical number used in my analyses. **Figure 3-2** shows the corresponding scatter for ln(P|D). Adequate burnin and MCMC cycle numbers are essential for reproducibility; in the literature there are cases (e.g. Hamblin *et al.* 2010; Neumann *et al.* (2010); Reimer *et al.* 2008) where it seems likely that insufficient cycles of model fitting have resulted in failure of STRUCTURE to stabilise with the potential to draw wholly misleading conclusions.

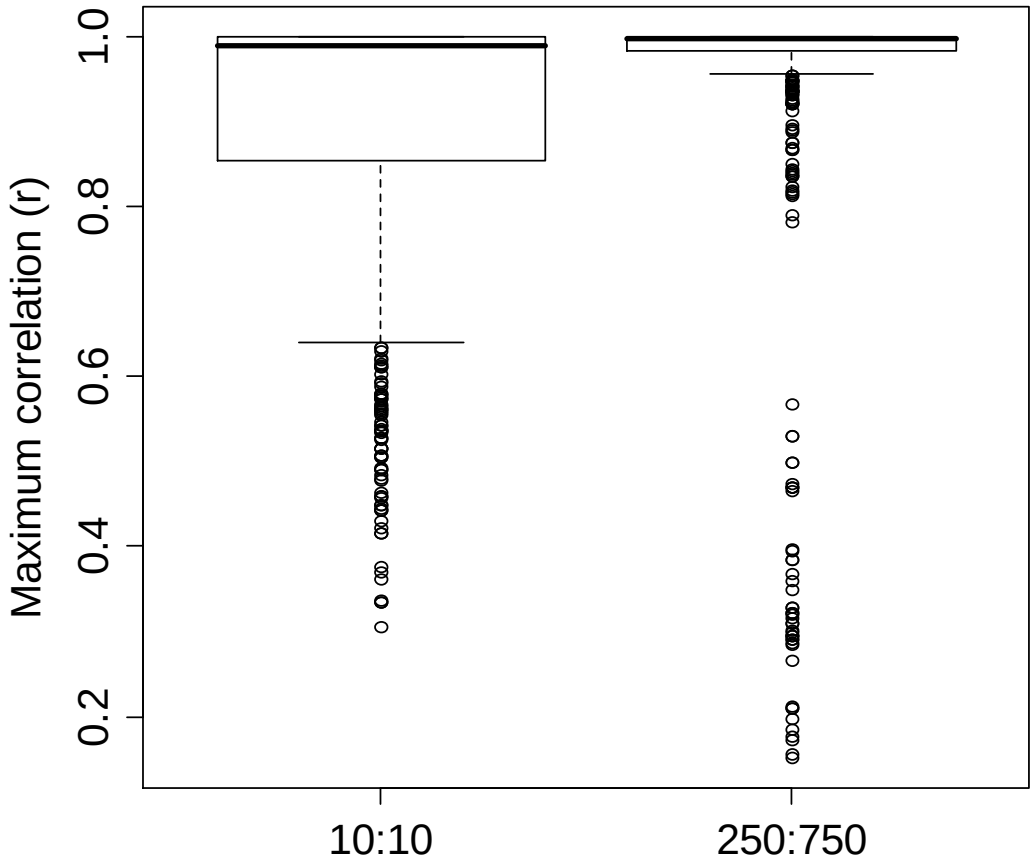


Figure 3-1. Box and whisker plot showing the scatter of maximum correlation coefficient between fractional sub-population allocations of varieties between 10 replicate runs with *k*=15 (all pair-wise comparisons). 10:10 indicates 1000 burnin and 10,000 MCMC; 250:750 250,000 burnin and 750,000 MCMC cycles.

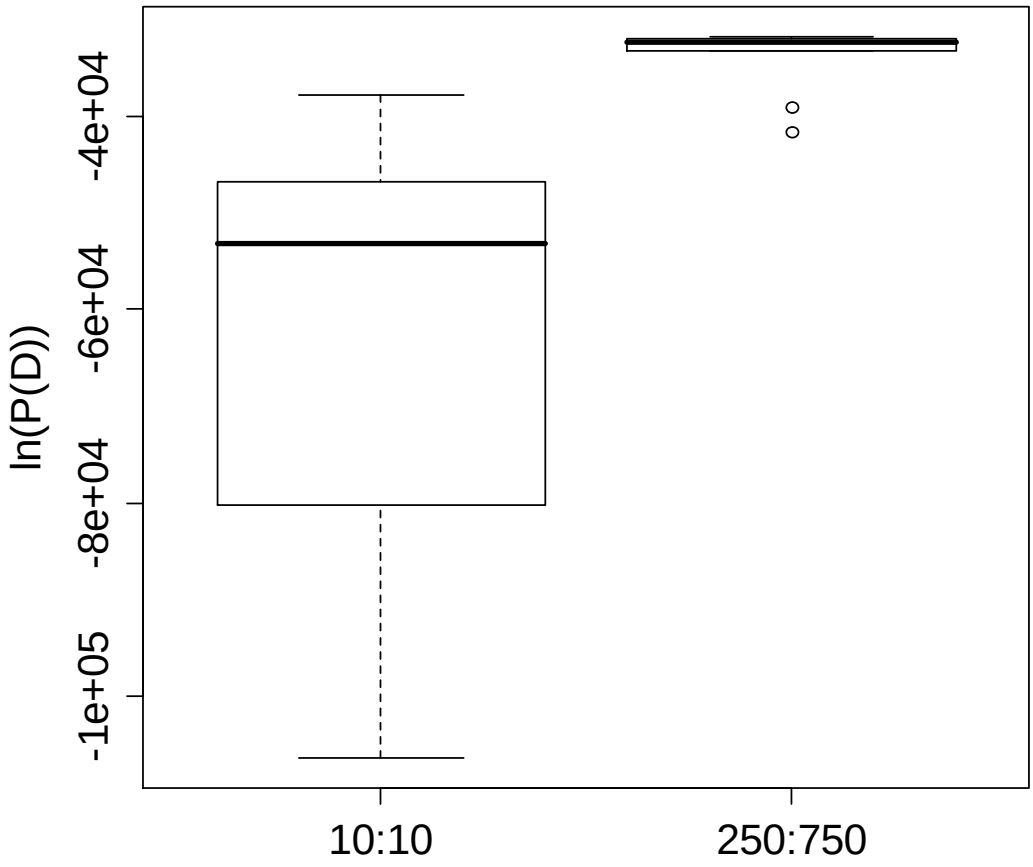


Figure 3-2. Box and whisker plot showing the scatter of ln(P(D)) of 10 structure models with *k*=15. 10:10 indicates 10,000 burnin and 10,000 MCMC; 250:750 250,000 burnin and 750,000 MCMC cycles.

3.3 Problems of kinship.

Whereas population structure is a potential source of spurious association measured at a macro level, kinship can have similar effects due to the micro-structure amongst close relatives.

In the context of association mapping, if two individuals are closely related it is likely, due to allele sharing, that they will share many aspects of their phenotype. Most of the alleles they share due to kinship will not be related to the specific trait under examination. The

consequence may be that alleles shared by reason of kinship will associate with phenotypic similarities. Like structure, kinship may result in spurious associations if its effects are not controlled.

Kinship may be estimated from pedigree data or from molecular markers.

3.3.1 Kinship based on pedigree.

The inbreeding coefficient (F) of an individual is the probability that a pair of alleles carried in the gametes that produced it were identical by descent (IBD) and is equal to the kinship (f) of its parents. In simple pedigrees:



If the kinships f_{AC}, f_{AD}, f_{BC} and f_{BD} are known then the kinship f_{PQ} is:

$f_{PQ} = ¼ (f_{AC} + f_{AD} + f_{BC} + f_{BD}) = K_{PQ}$

This formula gives the probability that any alleles taken at random from P and Q are IBD.

The use of pedigree in this way has been the method of choice for determining kinship for many years, it is unbiased but it also ignores the process of Mendelian sampling which leads to gradations of kinship between the progeny of a single cross.

The terms IBD and IBS (identical by state) need consideration. In the above we have the probability of IBD, without extensive genotyping of the entire pedigree we cannot specifically say which allele is IBD and which is not. Even if this were attempted it is reasonable to ask how far back should the pedigree go? If we take the progenitor lines as being unrelated then there is an internally consistent solution to K but this assumes the alleles present in the progeny arose independently even though they may be identical – they are IBS. A more honest view may be that, as we look back in time, the population has been mating randomly and chromosomes have been recombining; the alleles we see in the progenitors are

a random sample from that process. All alleles IBS may also be IBD; it is simply that recombination, Mendelian sampling and random mating have diluted the information about descent so that it is convenient to consider there to be two sources of identical alleles. The consequence of this compromise becomes more important as markers are used to estimate K.

3.3.2 Kinship from marker genotype data.

An alternative approach to the estimation of kinship is to use molecular markers to sample the genome of two individuals and look for similarity. This approach should take account of Mendelian sampling issues and also deal with hidden features of pedigree. For example, in the context of an inbreeding species such as wheat, and particularly when individuals have been developed as inbred lines through processes which include back-crossing and selection, and several generations of self pollination the picture is very complicated. In the pedigrees above we cannot tell if P has been repeatedly back-crossed into A with simultaneous selection for the introgression of desirable traits from B. Even if we knew this we would not know how many generations of back crossing were undertaken (each, on average, reducing f_{PB} by ½). Similar questions arise in relation to Q and f_{AC}, f_{AD}, f_{BC}, and f_{BD} are all similarly obscured as are the relationships of previous generations. This phenomenon is explored in detail by Maccaferri *et al*.(2007) in durum wheat accessions; they demonstrate the poor correlation between pedigree and molecular based estimates of kinship and also the effect of backcrossing as a source of bias between the approaches.

Given this uncertainty in the calculation of kinship from pedigree (and indeed the absence of pedigree data in many circumstances) an alternative approach is to consider similarity in genome as a measure of relatedness. This can be achieved in several ways:

Simple Matching Coefficient

In the simplest case, one might assume that each derived SNP allele arises from a single historic mutation event thus two individuals sharing a common derived allele have some small shared ancestry allowing a simple matching coefficient to be used in place of an estimate from pedigree (for examples see; Zhao, *et al.*, 2007; Kang, *et al.*, 2008):

$$K_{ij} = \frac{S_{ij}}{m}$$

Where S_{ij} = number of markers for which haploid individuals *i* and *j* share an allele, *m*= total number of markers. Estimates of *K* vary from 0 to 1.

In this case all shared alleles are assumed to indicate shared lineage – the alleles are identical by descent (IBD). However some alleles may be shared as a result of pedigree events so distant in the past that the sharing of alleles can more realistically be regarded as a result of chance (IBS).

Estimating kinship using a simple matching coefficient has the disadvantage that it ignores the distinction between IBS and IBD and hence gives equal weight to every marker. This will give different absolute estimates of kinship for the same individuals depending on the allele frequencies of the marker panel used. A simple illustration of this is given in Figure 3-3 which shows that although the genotypes have been assigned at random (which gives a panel of individuals who are averagely related to each other – that is they share alleles no more frequently than would be expected by chance) the simple allele sharing method gives different values for average relatedness dependent on the minor allele frequency of the marker panel. Low minor allele frequency results in many cases of individuals sharing the major allele and hence appearing highly related.

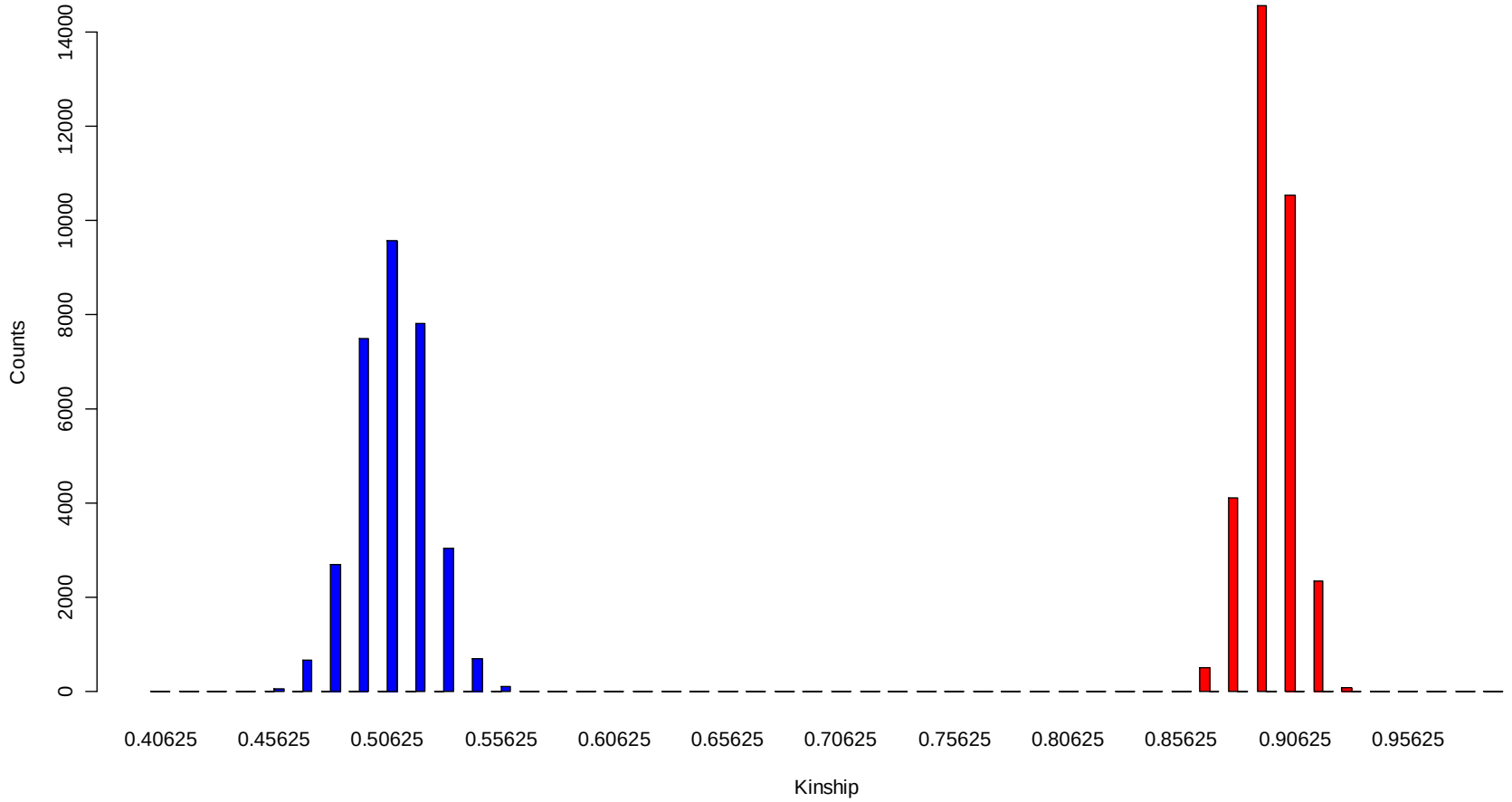


Figure 3-3. The distributions of pair-wise kinships estimated between 254 individuals assigned random genotypes at 1000 marker loci. Blue and red bars are estimates based on simple allele sharing when MAF was set to 0.4 - 0.5 and 0.01 - 0.1 respectively.

Excess allele sharing.

A better estimate of kinship from markers may be obtained if the alleles shared are notionally partitioned between those shared IBS and those shared IBD. In the absence of detailed pedigree data and genotypes of preceding generations it is impossible to make specific assignments for any marker in any pair of individuals but it is possible to estimate how frequently alleles at a locus would be expected to be shared between two individuals if each genotype is considered to be a random sample from a pool of alleles of known frequency. Thus an average relatedness can be estimated over all markers and deviation in the proportion of alleles shared, in either direction, can be viewed as a scale of relatedness; kinship.

It is important to note that IBD in the context of excess allele sharing is different from IBD estimated from pedigree; the pedigree records inbreeding and the estimate of IBD (in so far as the pedigree is known) includes this whereas using markers IBD is estimated as excess allele sharing under a hypothesis which is probably untrue (mating may not be at random, marker genotypes are not independent of each other, etc). Despite this many authors use the term IBD without reference to the specific assumptions behind it.

As Astle *et al*. (2009) explain, an estimate of kinship ($\hat{K}$) based on excess allele sharing is equivalent to an estimate of genetic correlation; in matrix form (diploid case):

$$\hat{K} = \frac{1}{L}\sum_{l=1}^{L}\frac{(x_l - 2p_l\underline{1})(x_l - 2p_l\underline{1})^T}{4p_l(1-p_l)} \qquad (1)$$

Where x_l is a vector of genotype (2, 1 or 0) over individuals for allele A of marker l, L is the number of loci, p_l is the allele frequency of allele A of marker l and $\underline{1}$ represents a vector the same length as x_l comprised entirely of 1's.

For the haploid case (applicable to inbred cereals as in this thesis):

$$\hat{K} = \frac{1}{L}\sum_{l=1}^{L}\frac{(x_l - p_l\underline{1})(x_l - p_l\underline{1})^T}{2p_l(1-p_l)}$$

Key to the estimate of excess allele sharing is an estimate of population allele frequency (*p*) at each locus. Given a panel of individuals a point estimate may be made from the data for each marker, however the resulting estimate of kinship will then be specific to the panel rather than the population.

A solution to this problem is offered by Lynch (1988) and Melchinger (1991). In this method a group of individuals, unrelated to the test population, are used to calculate S_{ij} – in this case, the average number of marker alleles shared between each control and the individuals in the test set. Kinship is then estimated as:

$$K_{Tij} = \frac{S_{ij} - 1}{1 - T} + 1$$

Where T is P(IBS|not(IBD)). The value of T is can be inferred by minimising an appropriate cost statistic (for example, in an association study of wheat varieties, Stich *et al.* (2008) adjust T to minimise mean squared difference between observed and expected P values (i.e. a P-P plot)). This approach allows the test panel, and any future test panel, to be given kinship estimates scaled to the same baseline.

Astle *et al.* (2009) suggest an alternative solution in which $\hat{p}_l$ is estimated from the initial estimate of $\hat{K}$ (1) and then iteratively each is re-estimated from the other until $\hat{K}$ stabilises using the relationship:

$$\hat{p}_l = \frac{\underline{1}^T \hat{K}^{-1} x_l}{\underline{1}^T \hat{K}^{-1} \underline{1}}$$

In the context of association mapping it may not be necessary to go beyond making an internally consistent estimate of kinship within the panel. The issue of allele frequency can be addressed by using the estimates from the panel. Given this simplification, then (1) may be re-stated as:

$$\hat{K} = \boldsymbol{X}\boldsymbol{X}^T / L \qquad (2)$$

Where $\boldsymbol{X}$ is a matrix of the panels genotypes (markers in columns) standardised to have mean zero and unit variance.

Returning to the example data used in Figure 3-3, Figure 3-4 shows the benefit of the use of the correlation matrix (as equation 2) compared to simple allele sharing. It can be seen that the disparity in estimate of average kinship caused by differences in MAF are removed when the estimate of kinship takes account of allele frequency.

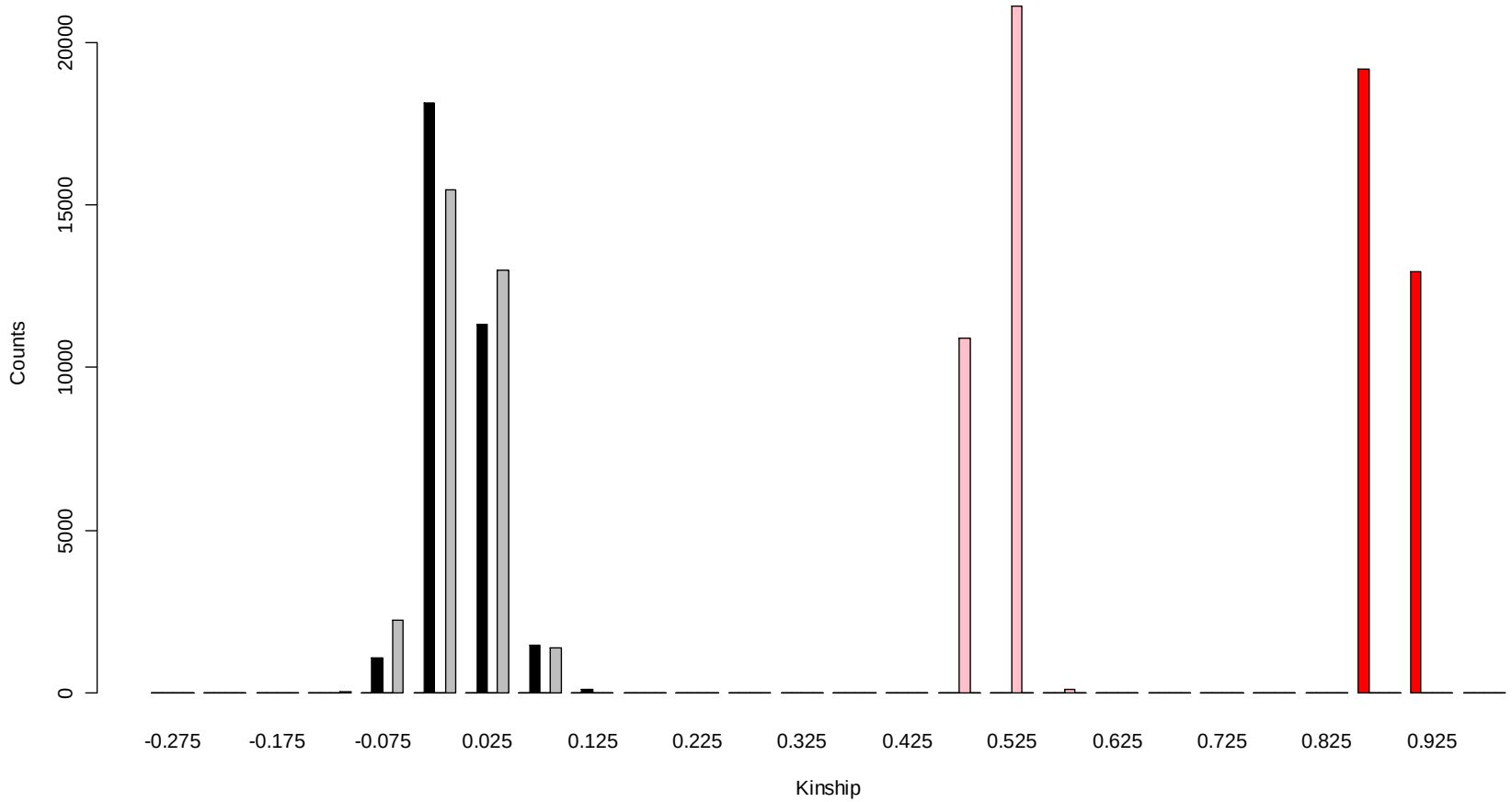


Figure 3-4. The distributions of pair-wise kinships estimated between 254 individuals assigned random genotypes at 1000 marker loci. MAF 0.01 - 0.1 red and black bars; MAF 0.4 – 0.5 pink and grey bars. Pink and red bars are estimates based on simple matching; grey and black are correlation based on standardised genotype.

Figure 3-5 shows the difference in estimates of kinship for the same individual using the two methods, the difference in ranking of pair-wise kinship is striking when MAF is less than 0.1.

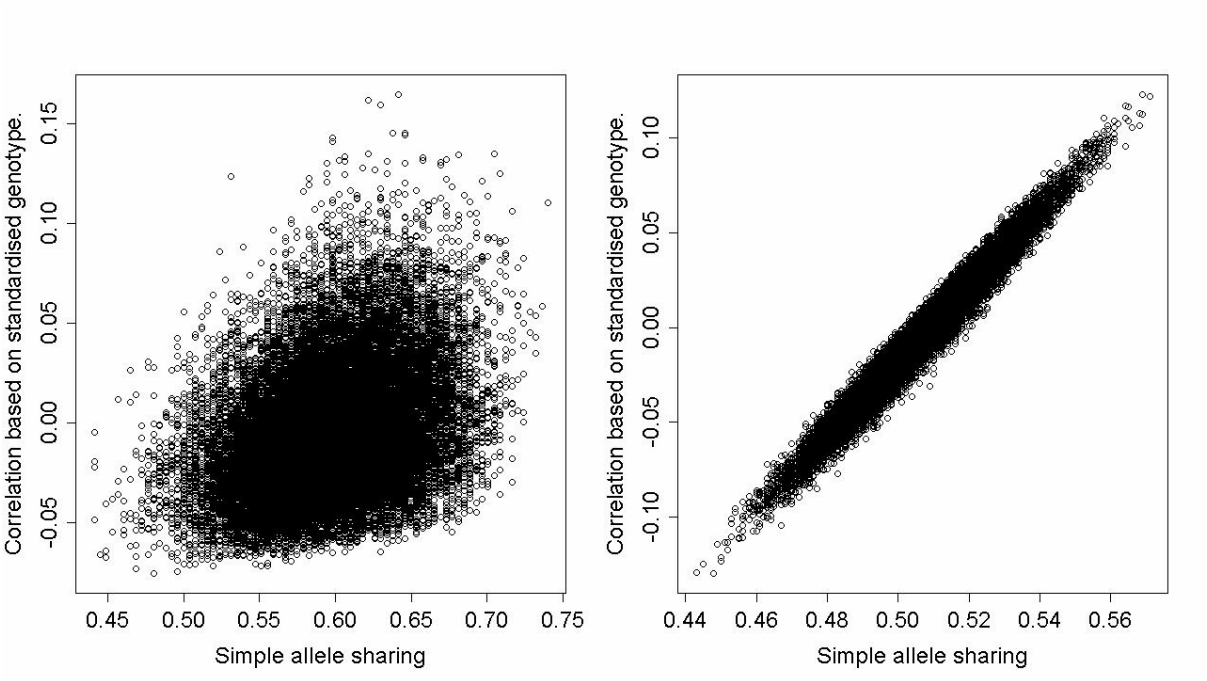


Figure 3-5. The relationship between kinship by simple allele sharing and correlation of standardised genotype. MAF is 0.01 - 0.1 (left hand figure) and 0.4 – 0.5 (right hand figure). When MAF is high there is little difference in the ranking of the estimates but at low allele frequency the two approaches diverge such that the ranking of pair-wise kinships is substantially changed in many cases.

3.3.3 Boundaries between kinship and population structure.

In the previous sections kinship and population structure were considered separately. This probably misrepresents the reality since both factors would be estimable from a complete pedigree.

If population structure can be regarded as a macro phenomenon affecting populations then kinship is a micro effect measured between individuals or families. The question as to where one ends and the other begins has no obvious answer and if both are manifestations of the same underlying processes then the question is almost meaningless. In this context it is interesting to explore the sub-population structure which the program STRUCTURE finds when only kinship is present.

Figure 3-6 shows the sub-population allocations for haploid individuals allocated random genotypes at 100 markers. A parental group was randomly mated to create 100 progeny who

were in turn randomly mated to create 100 grand-progeny. In the absence of sub-population structure the method finds structure within the recent kinship. Examination of the allocation of individuals to sub-population (when K=4) shows where individuals are found to have high fractional membership of a single population they tend to share one or more parents (Table 3-1).

Thus in this context major features of structure may reflect the breeding success of individuals (P25 in Figure 3-7).

The implication for populations of varieties is potentially very significant; in the absence of strong historic sub-population structure and in the presence of closely related individuals the **Q** matrix summarises major features of kinship. This is illustrated in **Figure 3-7** where, despite the random allocation of genotype and random mating, STRUCTURE is able to find groups of individuals based on successful breeders.

Table 3-1. The pedigrees of individuals allocated to a single sub-population by STRUCTURE. Parental lines have been designated P_x where x is a unique identifier.

Pedigree of individual.	Fractional membership of sub-population 3.
P25xP80	0.833
(P25xP27)x(P81xP77)	0.836
(P25xP80)x(P55xP30)	0.842
(P25xP80)x(P14xP94)	0.842
P13xP25	0.863
P25xP87	0.899
P25xP27	0.922
P25	0.928

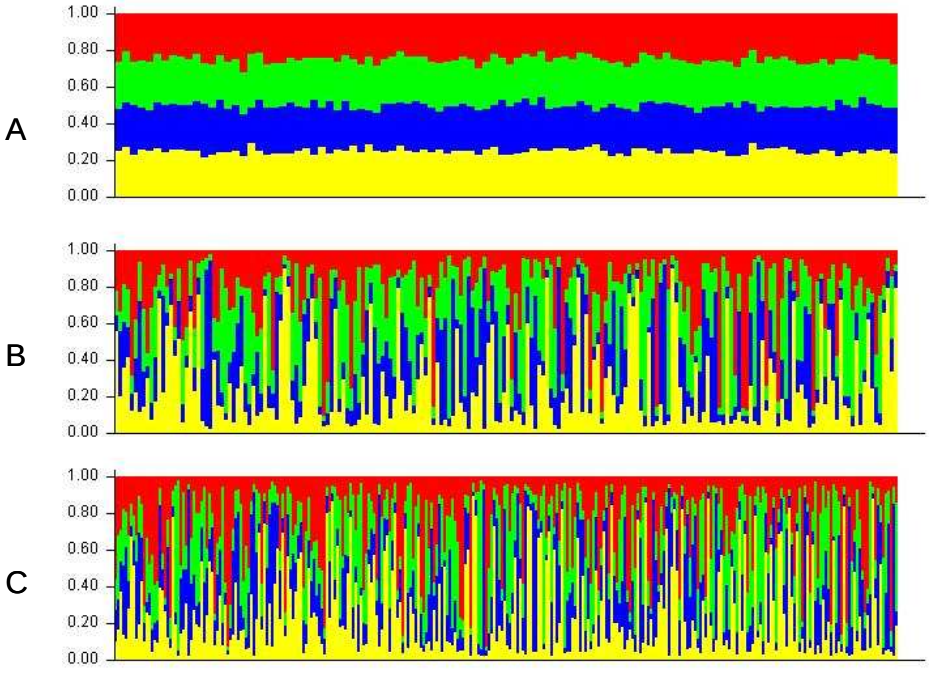


Figure 3-6. Assignment of individuals (vertical bars) to sub-population (colour) in the absence of structure. A. Parental population of 100 individuals, genotype randomly assigned for 100 markers. B. Parental population (as A, left half) and 100 progeny from random mating. C. Parent and progeny (as B, left two-thirds) and 100 grand-progeny from random mating within the progeny from B.

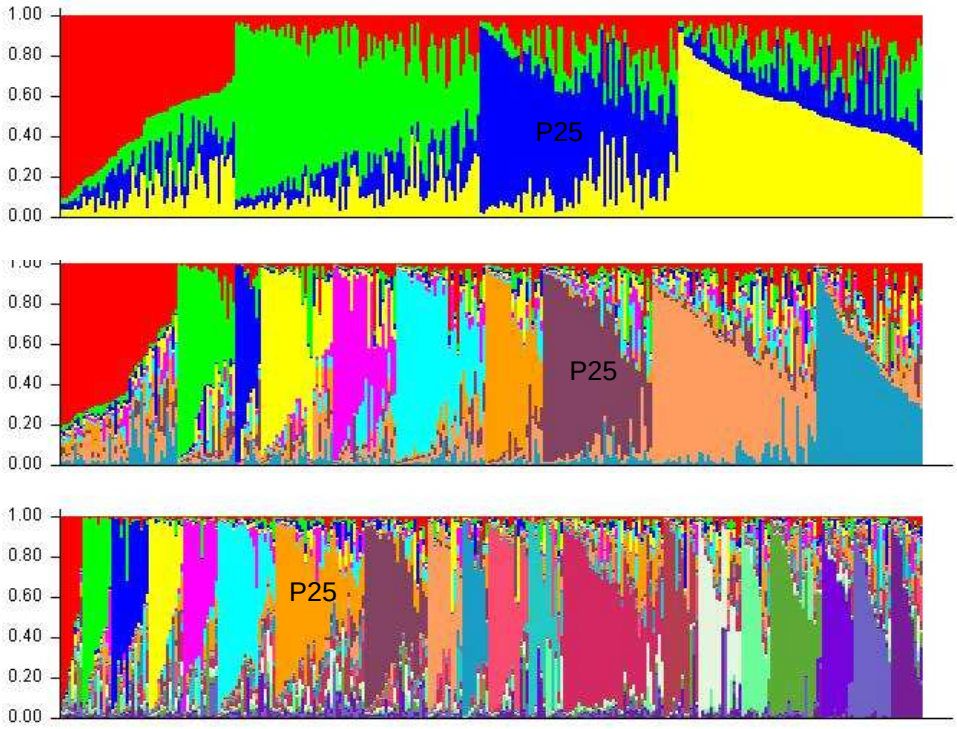


Figure 3-7. Assignment of individuals from Figure 3-6 C to 4 (top), 10 and 20 (bottom) sub-populations each of which has a family group as its nucleus. Highlighted in each case the group based around the P25 lineage (pedigree in Table 3-1). Individuals are sorted by population membership.

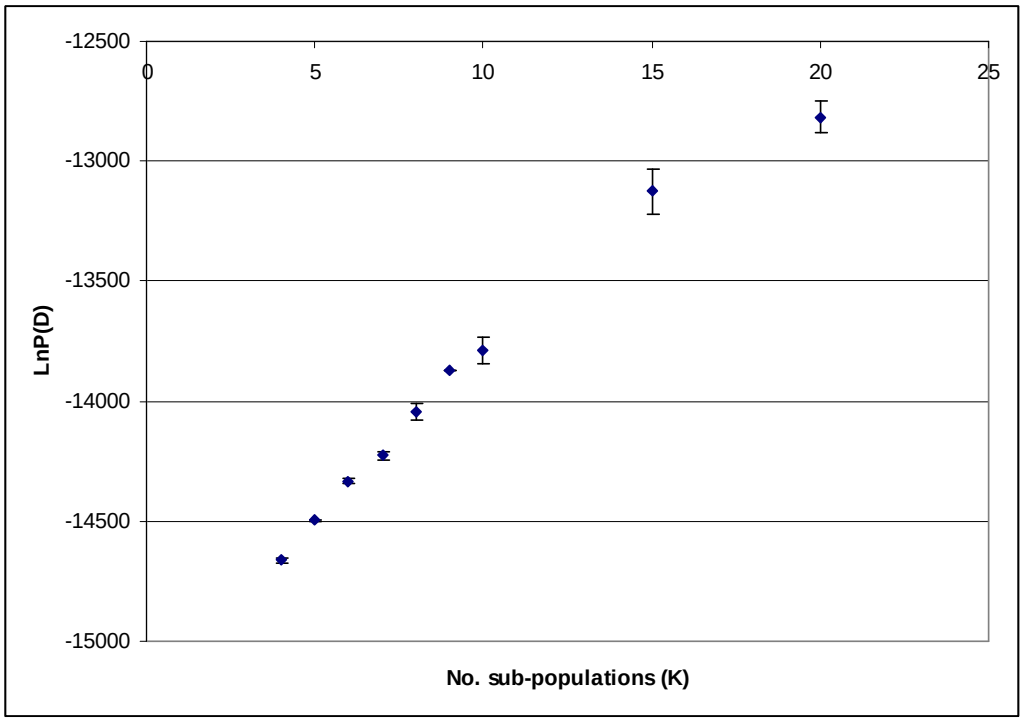


Figure 3-8. LnP(D) (± 1 SD) of models of sub-population structure derived from a simulated, unstructured and randomly mated population (Figure 3-6 C). (STRUCTURE output).

It is interesting to note that in these simulated data the performance statistic from STRUCTURE (lnP(D)) continues to rise as additional sub-populations are defined (Figure 3-8). This has been observed in experimental data in barley (Kraakman, *et al.* 2004) and was attributed to absence of structure but not directly associated with kinship.

Clearly kinship has the potential to manifest itself as sub-population structure in these small sample sets. The consensus emerging from association mapping studies of inbreeding species is that kinship should be measured and controlled for. This consistent with this small simulation. However, the observation that structure and kinship overlap suggests that any analysis including both kinship and sub-population membership may double count some information; a robust estimate of kinship may already include a model for structure.

3.3.4 Experimental design for the control of the effects of population structure and kinship.

In linkage (as opposed to association) studies it is possible to control for sources of spurious association by experimental design.

Classically the Transmission Disequilibrium Test (TDT) has been used in medical genetic epidemiology and uses family groups with heterozygous parents and affected offspring to test for excess transmission of one marker allele over another. (Spielman and Ewens, 1996). In plant science an equivalent technique would be the use of biparental mapping populations.

These techniques avoid the problems of structure; essentially they are their own control in this respect. However biparental mapping has the limitation that it can only be used to map traits and markers where there is measurable difference between the parental lines. The TDT test samples many family groups and so it samples a wide range of marker genotypes. However the panel is assembled with a specific phenotype in mind so the scope for using a TDT panel to map multiple phenotypes is limited. Both methods require large panels if the experiment is to have sufficient power to detect QTL of low effect.

Biparental mapping will be referred to in the course of the thesis since it is a complementary technique to association mapping.

3.3.5 Statistical control of the effects of population structure and kinship.

The effects of sub-population structure are genome wide, effects falling on all loci with varying strength. The effect of structure is therefore to inflate the value of the test statistic by varying amounts at all loci, the variation being due to differences in allele frequency between sub-populations. This effect can be used to demonstrate the presence and severity of its effects in a P-P plot of the observed log-probability distribution over the expected distribution under the null hypothesis (see Figure 3-9). Indeed general inflation of the test statistic gives the basis of the simplest method of statistical adjustment for the effects of structure; genomic control (GC) (Devlin and Roeder, 1999).

3.3.5.1 Genomic control (GC).

Under the null hypothesis the test statistic for association has a distribution characterised by, among other measures, a mean, median and variance. Inflation of the test statistic due to association between some markers and the trait results in a deviation from this expectation. Whether the trait is qualitative or quantitative the number of significant true associations observed in a genome-wide study is likely to be small meaning that the deviation of the observed distribution from the null is unlikely to be large under any credible genetic model. In practice, however, where spurious association is present it leads to widespread inflation of estimates of association across the whole genome.

The principle underlying GC is that although estimates of association will be generally inflated across the genome, the true associations will be among the most significant. The difficulty faced by the analyst is to know at what value of the distorted test statistic to declare significance. GC offers a solution to this dilemma by adjusting the observed distribution of the test statistic such that a chosen measure of centrality is returned to its null value.

Given that the observed values for the test statistic may not all come from the null distribution – there may be some true discoveries among them – then the choice of measure of centrality should not be excessively influenced by an extended tail on the distribution. Two measures of centrality have been used; initially Devlin and Roeder suggested the median. Taking a typical test statistic as an example, χ_1^2 has median ≈ 0.455. Inflation of the statistic might result in a higher observed median in the data. Dividing all observed test statistics by the ratio of the observed median to the null median returns a distribution of the test statistics centred as if it were from the null distribution. Subject to considerations of multiple testing (Chapter 3.5) determining significance thresholds becomes a standard procedure.

An alternative measure of centrality is the trimmed mean of the observed values of the test statistic (Clayton *et al.* 2005) in which a top slice of observed values are discarded (say the top 5 or 10%) before the observed mean is calculated.

Figure 3-9 shows the effect of GC in data where the test statistic has been inflated by structure or kinship. In this case genomic control (using the median) returns the distribution of the test statistic, and hence log (p-values), to the null but little evidence of significant association remains.

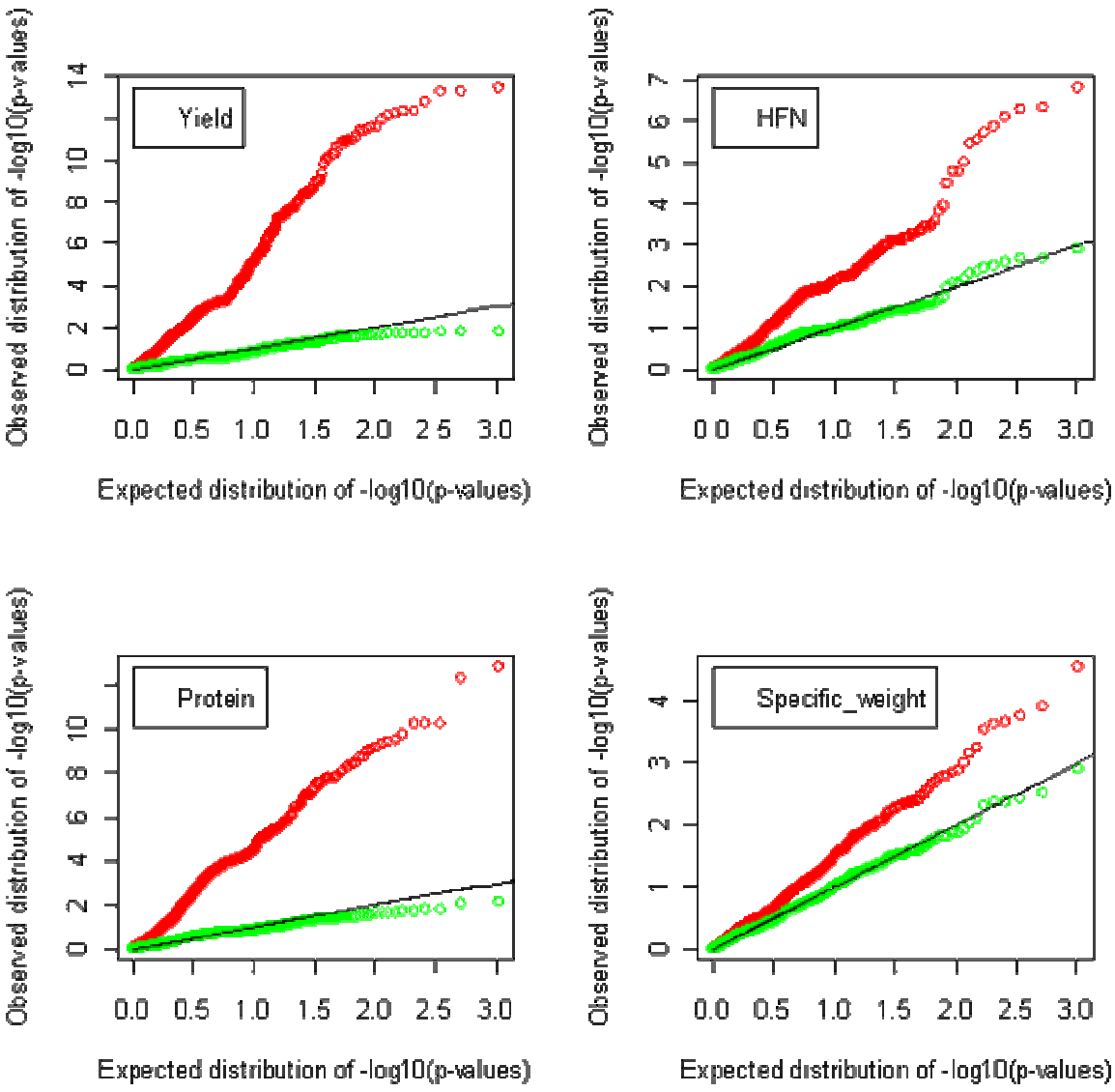


Figure 3-9. PP plots of naïve association (red) measured as correlation between marker genotype and phenotype and the same data after adjustment for inflation using genomic control (green). Panel comprises 167 wheat varieties genotyped with 424 DArT markers. The four phenotypes are discussed in Chapter 4.

A potential weakness of GC is that it assumes structure to have had equal effect on all loci. But since the effect is due to co-variation between allele frequency and differences in phenotype between sub-populations or over a range of kinship the spurious inflation must fall more strongly on loci whose allele frequencies co-vary with these factors. The lack of sophistication of the method has the strong potential to sacrifice power in order to control these false positives.

Whilst GC offers a way to measure distorted associations it probably only has merit as a method of adjustment when the signal from association is very strong compared to that from

structure. However, even if the data do not allow GC to be used as the primary method of control it is often the case that the more sophisticated regression models below will fail fully to control for spurious association. In this case GC can be used after regression to exploit the improved discrimination between true and false positives which the regression may have achieved. Figure 3-10 illustrates this effect.

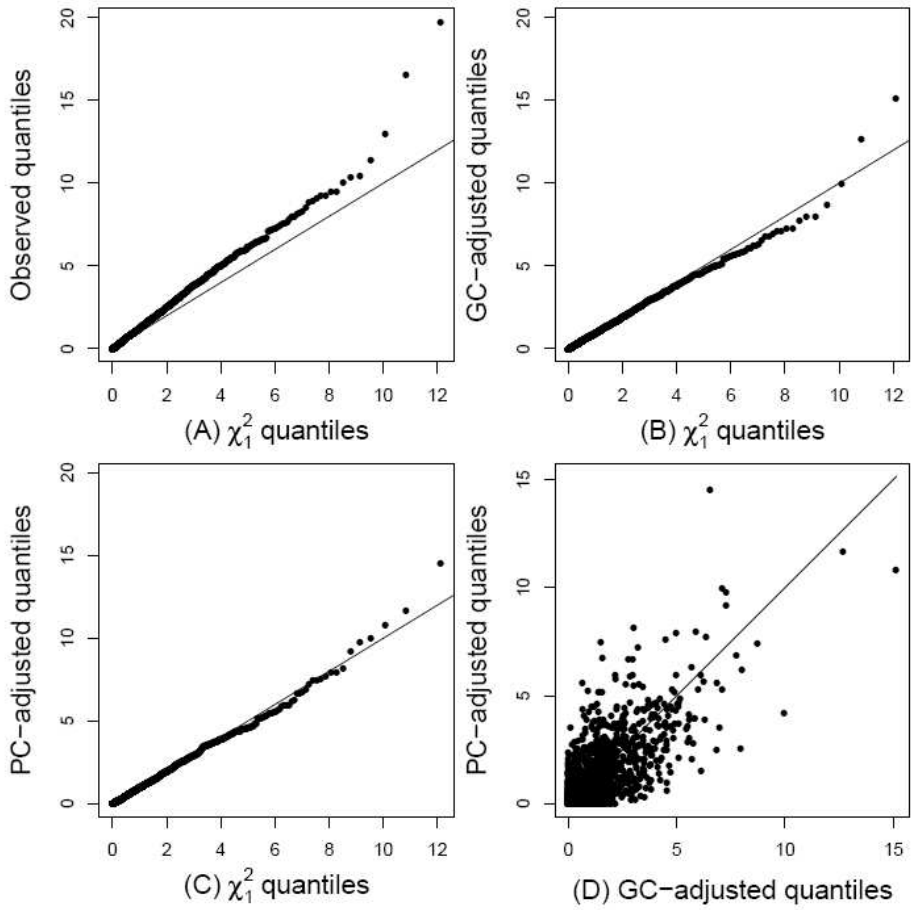


Figure 3-10. Q-Q plots for likelihood ratio tests of association in logistic regression in simulated populations (see; Astle *et al.* 2010 for detail). (A) shows genome-wide inflation of the test statistic, (B) shows the effect of correction using GC (median). By using GC to complement a regression model (C) in which the first two principal coordinates of a correlation matrix are used to model structure there appears to be a clearer distinction between the null (on the line) and the alternate (above the line) distributions. (D) shows that the regression model in (C) has substantially changed the ranking of the measures of association. (Figure reproduced from Astle *et.al* 2010).

Clearly GC has some merit but the erosion of power meant that more sophisticated approaches were needed for complex traits where individual marker effects may be low. More specific control of the effect of structure would have the potential to preserve power by selectively accounting for the association between markers and trait predicted by structure before measuring marker-trait association. A number of techniques have been developed to accomplish this; Structured Association (SA), Eigenstrat (PCA), Mixed effects modelling (MM) and Stepwise regression (SR). Respectively; Pritchard & Donnelly (2001), Price, *et al*.(2006), Yu, *et al.* (2006), Setakis *et al.* (2006). GC retains a role, however, because regression models may fail to completely control the effects of structure leaving some residual inflation of the statistic, this may be corrected with GC *after* regression modelling.

Association mapping using GC is readily achievable with modern statistical packages; I have implemented the method using R.

3.3.5.2 Structured association (SA).

SA was the first method designed specifically to allow sub-population specific control for structure (Pritchard *et al.* 2000a). In this method the sub-population model described in 3.2.1.2 summarised as the **Q** matrix is used as a covariate in regression of marker-trait association. Note that the rows of **Q** sum to unity so **Q** must be truncated by the removal of one column. Let **T** be **Q** truncated in this manner. Then the model for association is:

$$y_i = \alpha + T_i\ \gamma + x_{ij}\beta_j \qquad (3)$$

where y_i is the phenotype of the ith individual, T_i is a row vector comprising the fractional population memberships of the ith individual, γ is a column vector of regression coefficients for the columns of **T** (i.e the sub-populations), x_{ij} is the genotype of the ith individual at the jth SNP and β_j is the regression coefficient for the jth SNP.

The change in deviance between (3) and the equivalent model without the SNP approximates to a likelihood ratio test with χ_1^2 null distribution.

Association mapping using this method is readily achievable with modern statistical packages; I have implemented the method using R.

SA has been used in plant science [for example; Cockram, *et al.* 2008; Yu *et al.* 2006; Aranzana *et al.* 2005; Zhao, *et al.* 2007] but has three disadvantages. Firstly, there is an

inevitable loss of power due to over fitting as HWE and LD disequilibrium are minimised, secondly determining the appropriate value for *k* in any experiment is not straightforward (see;3.2.1.2) and thirdly the program STRUCTURE is computationally demanding such that large association studies cannot be analysed by this method. For example a robust implementation of SA may represent several weeks of PC processor time for even a modest dataset – 100-200 individuals with 200-400 markers.

3.3.5.3 Regression with principal coordinates as cofactors (PCA).

Conceptually simpler than SA, this method (proposed by Price *et al.* 2006) models the population structure using the first few eigenvectors (3.2.1.1) of a matrix of covariance between individuals calculated on normalised genotypes (the covariance matrix is thus closely related to the matrix of kinship calculated from excess allele sharing). In essence high levels of covariance would be expected to be associated with phenotypic similarity. Thus the expression of a trait, if associated with structure, should vary along the dimensions of the eigenvectors. Similarly, alleles whose frequency is associated with structure should also exhibit changing frequency along these dimensions.

Control is achieved by first using multiple regression to predict all the elements of the phenotype and genotype matrices from a matrix comprising the first *k* eigenvectors of the covariance matrix. The value of *k* (as in SA) is left to the researcher to determine; 20 is a default suggested by the authors but residual sum of squares of the prediction of phenotype might also be informative – as with the **Q** matrix (3.2.1.2, above). Alternatively the Tracy-Widom statistic has been suggested as a measure of statistical significance of the eigenvalues; Price *et al.* (2006) and Patterson *et al.* (2006) suggest including those vectors where the eigenvalues are significant at $P<0.05$ or $P<0.01$. In practice it would seem simplest to determine an appropriate number of eigenvectors for each phenotype in advance by predicting phenotype from increasing numbers of eigenvectors until the residual sum of squares is minimised.

Price, *et al.* propose a two-stage regression process, firstly genotype and phenotype are estimated using the first *k* eigenvectors:

$$y = \alpha + P\beta \qquad 4a$$

$$g_j = \alpha_j + P\gamma_j \qquad 4b$$

Where *y* and *g* are columns of phenotypes and genotypes of the panel, j indicates the jth SNP, **P** is a matrix of the principle *k* eigenvectors and, γ and β are column vectors of regression coefficients.

The second stage is to calculate the residual between $\hat{y}$ (from 4a) and the observed phenotype (y_r) and similarly g_{ri} from 4b.

SNP association with trait can then be tested as correlation of y_r and g_r; each SNP tested in turn with significance estimated from a t-test. The authors suggest that the false positive rate can be further controlled by the use of genomic control applied after eigenstrat provided a suitable one degree of freedom test statistic is used.

Association mapping using this method is readily achievable with modern statistical packages; I have implemented the method using R.

Use of PCA and its derivatives has been demonstrated in simulated data and is now being used in medical genetic epidemiology (Fellay *et al.* 2007. Hom, *et al.* 2008. Plenge, *et al.* 2007) and in plant science (Zhao *et al.* 2007 , Weber, *et.al*. 2007.).

3.3.5.4 Mixed effects modelling.

(A general description of mixed modeling is given in Appendix 1.)

The question, as to the relative contribution of kinship and structure to spurious association, has been considered by Yu *et al.* (2006) and Stich *et al.* (2008) and it is clear that the relative importance varies with the genetic history of the populations under study. In the case of inbreeding species it is evident that methods developed in human genetics may not be directly transferable.

In populations of plant varieties, groups of individuals will often share one popular breeding line in their recent pedigree; this stems from the desire to maintain useful genetic background intact and to only introgress small changes. Thus structure in plant populations might derive from recent breeding history more than historic population events.

Incorporating kinship into methods for control of structure has been attempted using MM (Yu, *et al.* 2006; Stich *et al.* 2008, Zhao *et al.* 2007). MM is a development of analysis of variance in which some of the factors are regarded as being samples from a larger population, the characteristics of which you are attempting to estimate (random effects) while other

factors are considered as fixed; the subjects of the experiment. Thus the association between kinship and phenotype is a random effect which is estimated from the sample population. The association between genotype and phenotype is a fixed effect. As with conventional analysis of variance the effects of each component of the variation are estimated in a linear model allowing the specific association of each marker on phenotype to be extracted from the total variation. Additional terms may be added to the model, and both Yu and Stich experiment with the addition of the **Q** matrix from Structure.

Figure 3-11 gives a summary of the results of Yu, *et al.* which show firstly that the effects of structure vary with the phenotype measured and secondly that kinship (in these inbred maize lines) gives better control of type II error than structure and that when structure and kinship are both included in the model the improvement over kinship alone is slight.

In Cockram *et al.* 2010 I illustrate the effect of structure/kinship and the power of the MM as a method of selectively controlling its effect (**Figure 3-12**). The striking feature of the figure is the selectivity with which off-chromosome and long-distance associations, widespread following regression, are removed by the MM. Note the almost complete lack of significant association above 20cM in the right-hand plot.

It seems likely that for inbred plant species, and particularly populations of elite varieties, control of structure using kinship should be the method of first resort. The software most commonly cited for the implementation of MM in this context is TASSEL (http://www.maizegenetics.net/index.php?page=bioinformatics/index.html)). However TASSEL is designed as an analytical (rather than research) tool. A single analysis of a dataset set is relatively easy to achieve albeit slow compared with available alternatives. Where multiple analyses of the same data with slight perturbations are required (for example in the estimation of statistical power) TASSEL lacks speed and flexibility and I have used the Efficient Mixed Model Association (EMMA; Kang et al., 2008) a rapid implementation of which is available from William Astle at Imperial College (http://astle.net/wja/) which has been my method of choice for association mapping using MM.

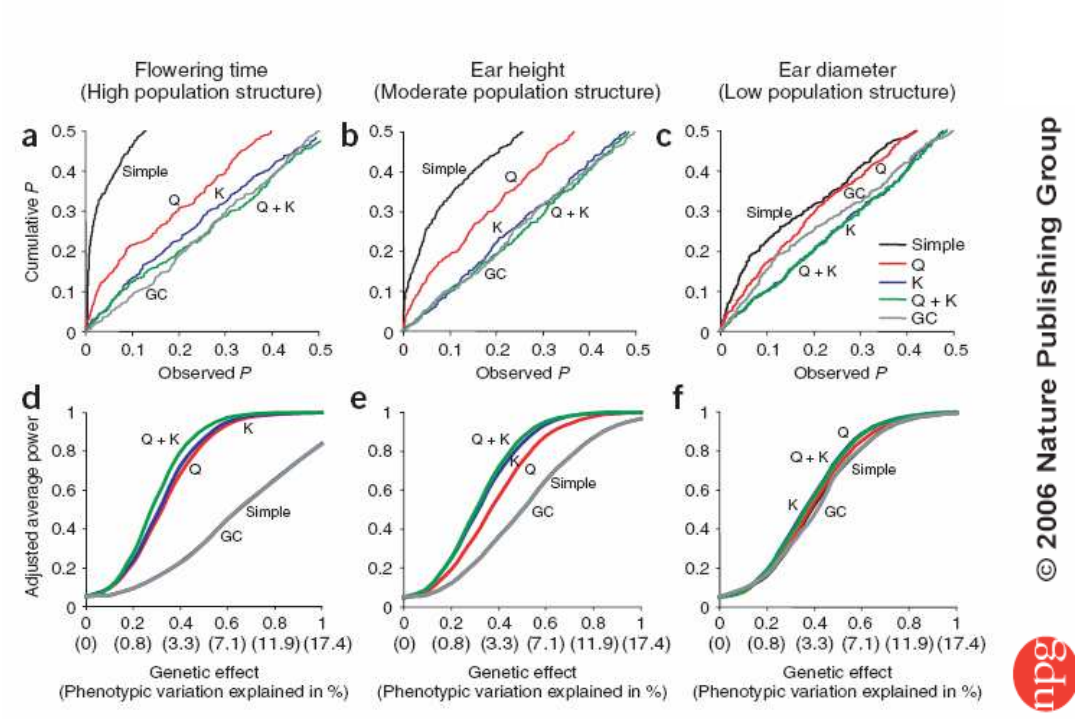


Figure 3-11. Power curves and cumulative probability plots showing the effect of controlling for structure with genomic control (GC), Sub-population membership matrix from STRUCTURE (Q), Mixed Model using kinship (K) and Mixed Model using K and Q (K+Q). The uncorrected (Simple) results are also shown. Maize inbred lines. (Reproduced from; Yu *et al*., 2006).

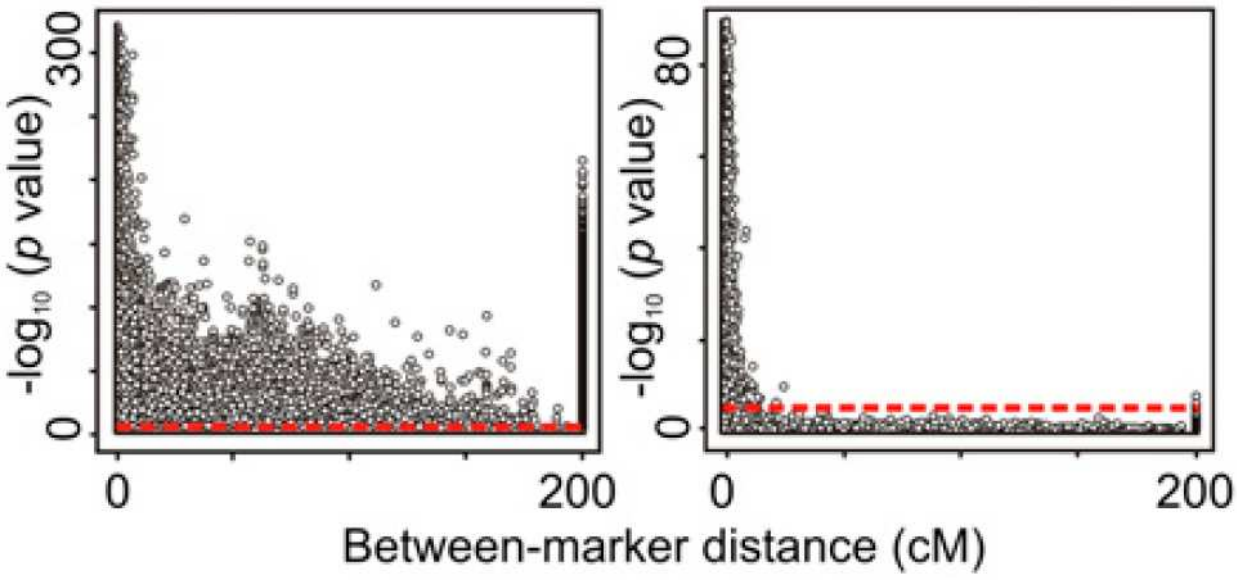


Figure 3-12. Between marker pair association over between marker pair genetic distance in the barley panel (described in 5.1). The left hand plot is association measured by regression the right hand plot are the same comparisons with association measured with the MM. All off-chromosome comparisons ar plotted at 200cM. The horizontal line is the Bonferroni correct P=0.05 threshold. (Reproduced from Cockram *et al.* 2010.)

3.3.5.5 Using null markers as co-variates in logistic regression

SA, PCA correction and the MM all attempt to control for the effects of structure by measuring or modelling structure/kinship and then applying corrections tailored to each marker in the panel. There is no doubt that each of these approaches will reduce the Type I error rates and so give more reliable detection of meaningful association than is possible from raw data or with genomic control. Two contrasting approaches, which make no attempt to model or measure structure and yet appear able to give similar control of error are Stepwise Logistic Regression (SLR) and Bayesian Logistic Regression (BLR) (Setakis et al., 2006).

Stepwise regression (SR) is a technique used to select a sub-set of explanatory variables from those available using an iterative process in which variables are either added to a base model, removed from a full model or a combination of the two. Addition or removal is determined by F or t tests, r^2 or information criteria. The goal of the iterative process is to find a stable regression model which minimises the residual sum of squares. Models which are parsimonious in use of predictive variables tend to be more stable when extrapolated beyond the experimental population.

In the method proposed by Setakis *et al.* backward SR is used from an initial model including all available markers. Inclusion and removal are considered for all markers at each step and a penalty (multiple of the degrees of freedom) of 4 ln(n) (where n= No Markers) is used. This penalty answers, to some degree, the criticism often directed at SR that the model is being both constructed and tested on the same data leading to considerable risk of over-fitting.

On simulated data Setakis shows SLR and BLR control false positive rates as well as SA. It is surprising that so simple a technique should give results comparable to those obtained by the more sophisticated methods. The authors suggest that the explanation for its success lies in the iterative process: When all markers are included in the model the effects of structure are present in multiple loci none of which is significant on its own. Thus during the early phase of the stepwise process structure is modelled, although not explicitly. When the final regression model is found, structure is no longer modelled within it but the remaining significant markers tend to be causal or at least proximate to the causal SNP. This is not to say that the model has chosen the most strongly associated SNPs but rather that only SNPs in genuine association consistently survive the early iterations of the stepwise process when structural effects are being implicitly modelled.

SLR and BLR offer considerable advantages over SA in terms of computational effort.

3.4 Performing a test for association.

Depending of the nature of the data and the statistical control deployed the null hypothesis of no association between marker and trait can be tested in a number of ways, the more frequently used methods are summarised in Table 3-2.

Table 3-2. Commonly used tests for association.

Phenotype.	Genotype.	Statistical control for structure.	Statistical test.	Specific null hypothesis.
Binary	Binary	None	Armitage trend test	Regression coefficient = 0
Binary	Binary/ Multi-alleleic	None or Q-matrix from STRUCTURE.	Likelihood ratio test on logistic regression model.	The addition of marker genotype as an explanatory variable does not explain any more variation in the response variable than the base model.
Continuous	Binary/ Multi-alleleic	None or **Q** matrix from STRUCTURE	Multiple Regression: F-test.	The addition of marker genotype as an explanatory variable does not explain any more variation in the response variable than the base model.
Continuous (following adjustment for structure)	Continuous (following adjustment for structure)	Principal component analysis (e.g. eigenstrat)	t-test for the significance of correlation between adjusted phenotype and adjusted genotype.	Correlation coefficient is zero.
Continuous or binary.	Binary/ Multi-allelic	MM including kinship and/or structure.	F-test, Wald test, likelihood ratio test.	The addition of marker genotype as an explanatory variable does not explain any more variation in the response variable than the base model.

3.5 Statistical Significance and multiple testing.

The threshold for statisistical significance in association studies is complicated by multiple testing. This refers to the large number of parallel single tests for significance which take place; if each of these in turn is evaluated against a threshold – for example if nominal α =0.05 then the expected number of false rejections of the null hypothesis over *n* markers will be $n\alpha$. Where *n* is large discoveries made against an unmodified value of α will be numerous and predominantly spurious. One solution, therefore, is to reduce the value of α.

Where the parallel tests are independent of each other a suitable adjustment to α gives the experiment-wise α (which I will designate, α') this can be calculated from the binomial distribution and approximates to $\alpha'=\alpha/n$ where *n* is large (this is the Bonferroni correction). However, since the assumption of independence is unlikely to be true for markers in LD this correction is too conservative and potentially leads to high type II error rates.

Westfall and Young (1993) (generically) and Churchill and Doerge (1994) (in the context of genetics) suggest simple permutation tests as a way to derive an appropriate value for α'. In the Churchill and Doerge method the phenotypes are permuted relative to the genotypes in the panel and association by marker is tested for each permutation. For each permutation the highest significance across all markers is noted and a frequency distribution of these significances is used to determine a revised significance threshold, for example the upper 95th percentile would give a revised P=0.05 threshold. Depending on *n*, α and, to some extent computing capacity, at least 1000 and preferably more than 10000 permutations are required to establish a value for α'.

Permutation has two important advantages over the Bonferroni correction; (a) it gives a valid significance threshold even where data depart from an assumed probability distribution (b) it provides a significance threshold specific to the experimental panel. However, association analysis with large panels is computationally intensive, to repeat the analyses 10,000 or more times to establish a significance threshold would become an unsupportable overhead in many situations. An alternative solution is to use false discovery rate (or q-value).

3.5.1 Using false discovery rates to determine significance.

Rather than extensively sample the null distribution by permutation or make an unjustified assumption about independence it can be useful simply to estimate the fraction of results likely to come from the alternate distribution rather than the null. This is the basis of the false

discovery rate (FDR); the proportion of falsely rejected hypotheses amongst all rejected hypotheses, which was developed by Benjamini and Hochberg (1995). Using FDR instead of α' allows the experimenter to; control the number of candidate markers needing confirmatory analysis, quantify the expected number of true associations amongst those markers and make comparisons between strategies deployed to control for structure.

Storey (2002;2003) further developed the FDR metric and Storey and Tibshirani (2003) give relatively straightforward guidance for its application. To summarise; consider Figure 3-13 in which two distributions are plotted. The null distribution is uniform, the alternate distribution is skew towards P=0. At any p-value the results declared significant will comprise a mixture of true (purple) and false (blue) discoveries, the ratio of the false to the (false + true) is the false discovery rate. It is evident that as p-values get smaller so too does FDR.

The **FDR at any stated p-value is the q-value**; the relationship between p-value and q-value is specific to the experiment and so gives a convenient solution to the problem of multiple testing.

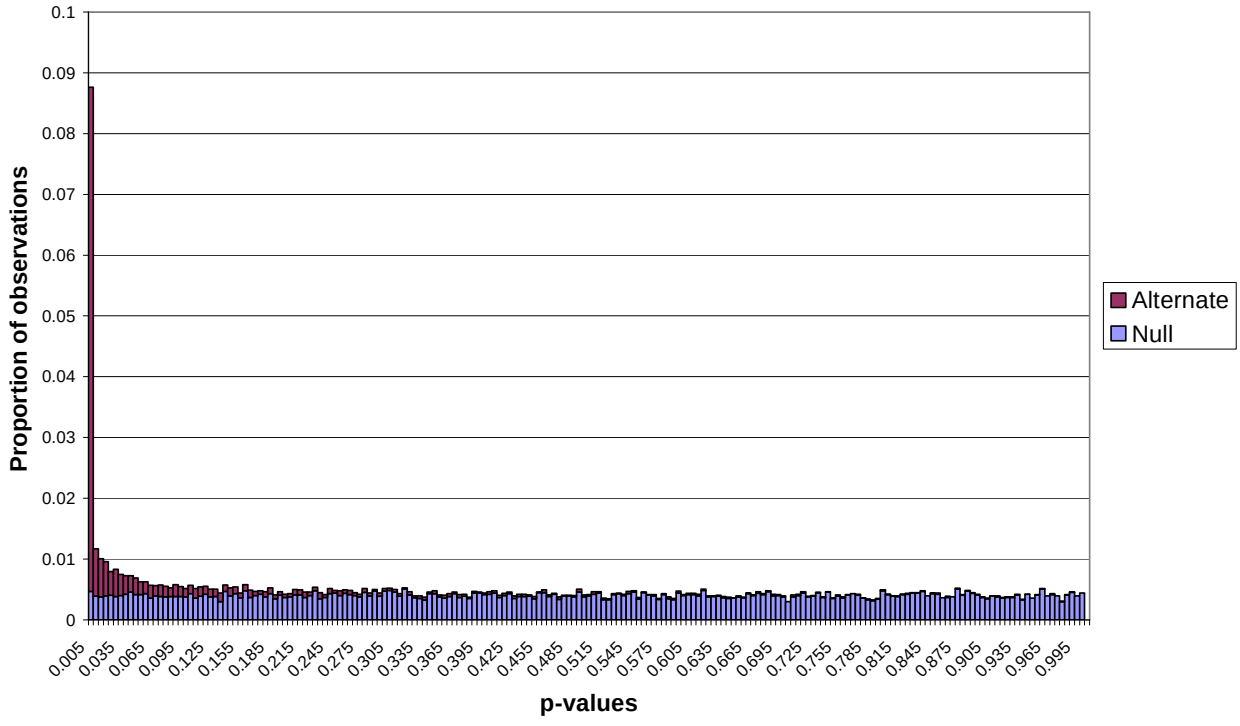


Figure 3-13. Histogram of p-values computed under the null χ_1^2 distribution for (blue) $2x10^4$ random observations sampled under the null; and (purple bars) $5x10^3$ random observations taken from an alternate distribution (χ_1^2 scaled by a factor of 8); the experiment-wide FDR for these data is therefore 80%.

From Figure 3-13 it is evident that if the number of results coming from the alternative rather than the null distribution were known, then the q-value at any p-value could be calculated. In practice, the calculation of FDR in real data relies on an assumption that above a threshold λ, all observed p-values come from the null distribution. If the tuning parameter λ were known then the null distribution could be approximated and q-values calculated. Since a proportion (1- λ) of null p-values is expected to exceed λ, an estimate of the number of null results is given by:

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1,..., m\}}{m(1-\lambda)}$$

where π_0 is the number of observations from the null (m_0) divided by the total number of observations (m), with the observed p-values in ascending order ($p_1 \ldots p_m$)
$\#\{p_i > \lambda; i = 1,..., m\}$ is the number of p-values greater than λ. Figure 3-14 illustrates how $\hat{\pi}_0(\lambda)$ varies with λ: the estimates should be less biased as λ increases, but based on fewer observations and so their variance increases.

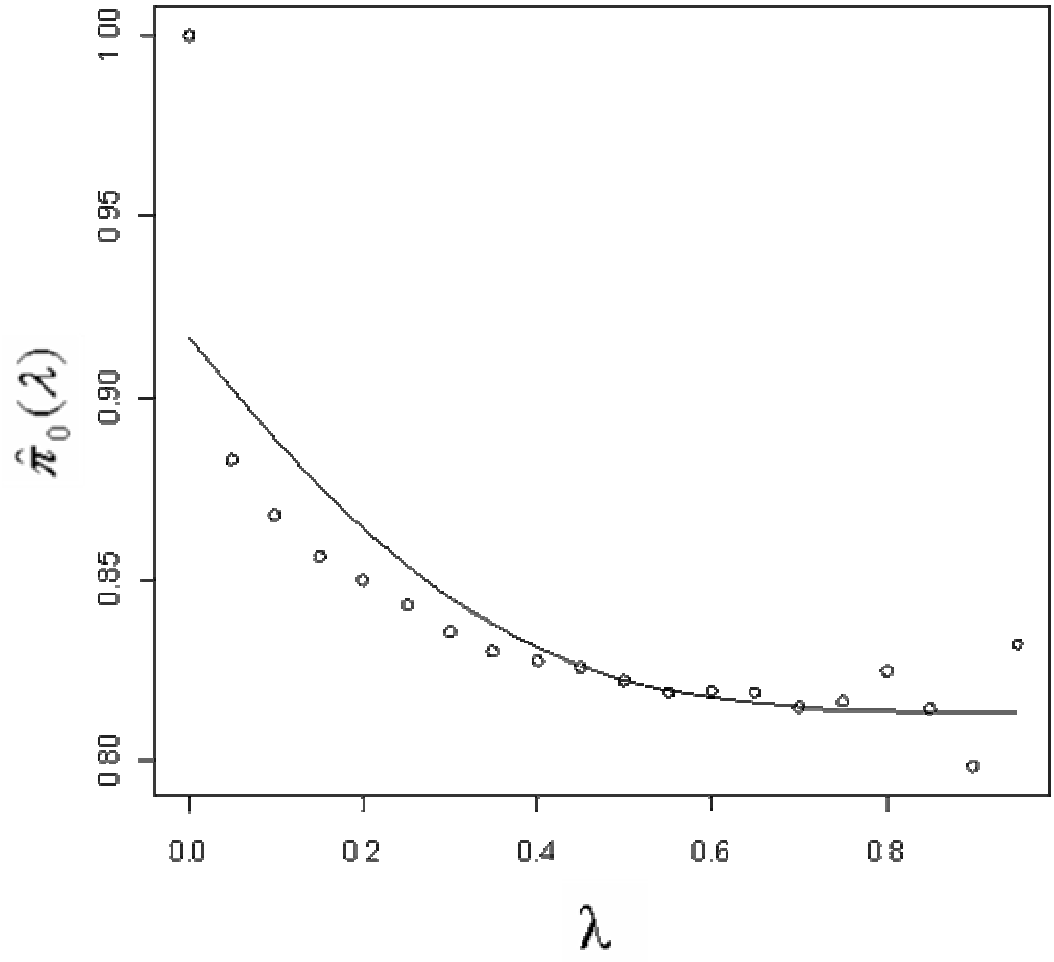


Figure 3-14. Estimates of $\hat{\pi}_0(\lambda)$, the proportion of p-values generated under the null, as a function of λ, the tuning parameter. Fitted line is a cubic spline (3df), the poor fit for low values is acceptable since larger values of λ give less biased estimates. The underlying data are those used in Figure 3-13.

One might look for an asymptote in the fitted line and interpolate experiment-wide FDR at that point, but since the proportion of observations from the alternate will be lowest at highest p-value the least-biased estimate of $\hat{\pi}_0(\lambda)$ is at λ close to 1. In Figure 3-13 this estimate of FDR would be ~0.82 (c.f. the true value is 0.8). It should be noted that estimates made in this way tend to be conservative, they err towards false rejection.

Given an estimate of $\hat{\pi}_0$, it is possible to estimate the q-value at any given p-value:

$$\hat{q}(t) = \frac{tm\pi_0}{1 + (\#\{p_i < t\})}$$

where t is the p-value of interest and $\#\{p_i < t\}$ is the number of observed p-values less than t.

In conclusion, q-values give a conservative estimate of the FDR rate at any given p-value. The estimate is specifically tailored to the data-set and so incorporates the number of tests made together with any lack of independence among those tests. Thus the q-value gives an easily accessible solution to the issue of declaring significance in association mapping experiments. An experimenter may choose a q-value of 0.1 below which all results will be declared significant. This may result in 100 markers appearing to be associated with the trait. If resources permit confirmatory work on these 100 markers then the q-value warns the experimenter to expect about 10% of the markers to have been false discoveries. If resources only permit further work on 10 markers the experimenter could reduce the q-value accordingly, perhaps to q=0.005, he would declare 10 markers significant and it would be unlikely that any came from the null. Note that this reduction in FDR has been achieved by discarding 80 discoveries which were from the alternate distribution (100-10-10=80).

There is one final point to make about FDR; the term 'false discovery' means a discovery attributed to the alternate which should have been attributed to the null. Linguistically the opposite of false is true, however in this case discoveries not attributed to the null are only true in the sense that they are unlikely to belong under the null; FDR does not offer any guidance as to which of the multiplicity of potential alternate hypotheses the discoveries belong.

3.6 Statistical Power.

I have outlined a number of analytical approaches and described how a suitable significance threshold might be found in an association mapping experiment. There remains the question of how to recognise the best approach overall. The answer will centre on the question of statistical power and I devote Chapter 6 to the question. However, it is appropriate to outline the principles involved at this point.

Statistical power is a measure of discrimination between two competing hypotheses: (a) the null hypothesis that there is no observable effect due to the treatment in the experiment and (b) the alternate hypothesis that the observed result is such an unlikely outcome under the null hypothesis that another explanation must be sought.

Power is often explained in terms of sources of error: Given an experiment with a statistical significance threshold, α, then an observation will be assigned to the null if the probability of making this observation under the null hypothesis is greater than α or otherwise it will be

assigned to the alternate. This highlights two sources of error. Type I error; α is the proportion of results from the null distribution that are assigned to the alternate: **false rejection of the null**. If there is a real treatment effect under the experimental conditions pertaining then the test statistic will correspond to p-values skewed towards low values when calculated under the null. Thus a higher proportion of p-values will be below α than under the null, but some and perhaps many of the p-values will exceed α and so will be accepted: **false acceptance of the null,** known as Type II error and its probability for a given alternate hypothesis is denoted β. The power for a given alternate hypothesis is 1-β.

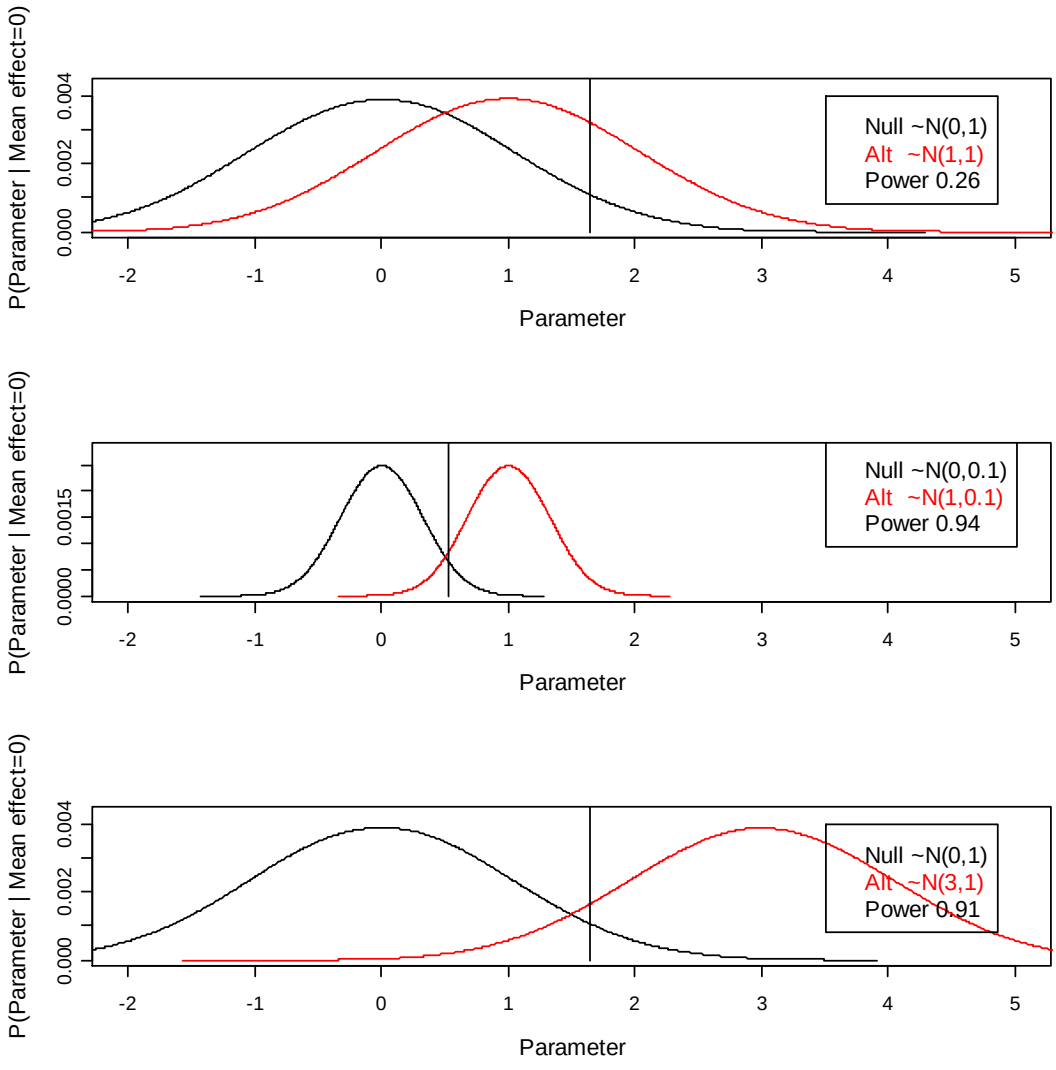


Figure 3-15. Statistical power, the ability to discriminate null (black) from alternate (red) distributions, varies with the value of the measured parameter and the precision with which it is measured. Vertical line is the upper P=0.05 significance threshold for rejecting the null hypothesis (α).

As the effect size specified by the alternate hypothesis rises the power will increase – intuitively it is easier to detect large effects than small effects – but also the precision with which the effect is measured will influence power because even a small effect measured with high precision will have a sampling distribution that overlaps little with the null distribution. I attempt to illustrate this in Figure 3-15 where the null hypothesis is that the treatment has no effect on the measured parameter and the alternate is that the treatment increases the value of the measured parameter. Note that high power (>90%) can be achieved if the effect size is large or if the variance is low

In the context of association mapping this allows the identification of factors which should influence the power of any experiment:

- Effect size of a QTL: Large effect sizes will be easier to detect; Mendelian traits may be easier than quantitative traits.
- Heritability: Traits which are highly heritable have been measured in a way which minimises the unaccounted environmental variation, this increases the precision with which genetic effects are estimated.
- Panel size: Large panels allow an effect to be sampled extensively reducing the variance of the sample distribution and increasing precision.
- Fit of regression model: Models which accurately incorporate the effects of structure/kinship (etc) with the effect of markers and discriminate between the two will minimise the residual error and increase precision.

A prior consideration of power can inform experimental design. After an experiment is complete prior estimates of power can be used to identify experiments in which the results are 'too good' and should lead the analyst to re-examine the experimental design, the statistical analysis and, indeed, the initial estimate of power to try to reconcile the difference. When the experimental design is held constant but the statistical methodology is changed power provides a useful metric for comparing methodologies.

Power calculations are not reported in the plant association studies literature as often as they should be, I note three instances, one of which is my own work Yu *et al*. 2006 (see Figure 3-11) use power to determine the best analytical approach for mapping in a population of inbred maize lines. Zhao *et al*. 2007 in Arabidopsis and Cockram *et al.* (2010) used *in silico* estimates of power to estimate the likely success rate for detecting botanical characters in

barley; the prior estimate is broadly consistent with the final result from association mapping over multiple traits.

4 Association mapping in plant species: Literature review.

In plant genetics the available marker density is still two to three orders of magnitude lower than that in the most recent human studies Nonetheless the numbers of markers are increasing and there have recently been a number of publications in which the application of association genetics approaches have been reported.

4.1.1 Barley (self pollinating diploid).

As early as 2004 Kraakman, *et al.* reported a study of 146 two-row spring[1] barley varieties from NW Europe using 236 mapped AFLP markers. The authors used STRUCTURE, genetic distance based phylogenetic trees and correspondence analysis (a generalisation of PCA allowing the use of categorical variables) to try to identify population structure. The correspondence analysis and the trees suggested two groups within the population but the authors could not find a reason for the division in the known attributes of the samples (e.g. malting type vs. feed type). In the case of STRUCTURE they report a failure of lnP(D) (the likelihood of the model) to stabilise with increasing K. Taking these outcomes together they conclude absence of structure.

It is possible that the test population may genuinely not contain significant historic sub-population structure, however the author's observation of continually rising lnP(D) and failure to allocate varieties unambiguously to sub-population, are not in themselves convincing evidence of this. As reported above (3.3.3) the boundary between kinship and structure may be unclear and this observation (of rising lnP(D)) may be evidence of STRUCTURE finding kinship groups in addition to sub-population structure; with each increase in *k* as a new kinship group is allocated a sub-population. The authors separately consider kinship as a confounding factor but do not pursue the question, mainly because of incomplete pedigree data.

A possibly better test for the presence of structure would have been to look for inflation of the test statistic over null markers. A pragmatic estimate of the relevance of any structure

[1] 2-row denotes the structure of the seed head, in this case containing 2 rows of grains. The alternative is 6-row, containing six rows. Winter varieties are sown in autumn and have a physiological requirement for a period of about four weeks of cold weather before flowering is initiated. Spring varieties are usually sown in spring and do not have this physiological need for cold weather.

observed might have been estimated simply by predicting phenotype from the terms in each of the **Q** matrices in turn. The residual sum of squares would show whether adding more sub-populations to the model was offering improved control of potential spurious association.

It would also be useful to know how many cycles of burnin and MCMC were used in the STRUCTURE runs since instability in lnP(D) can be a result of insufficient runtime.

The authors mapped their markers using mapping populations and demonstrated LD decay over a distance of about 10cM (measured as R^2 between chromosome marker pairs). The distance over which LD can be detected is quite high when compared to maize (2000bp) and sugar beet (3cM), the difference is attributed to the inbreeding nature of the crop and the effects of selection.

Association was measured marker by marker as correlation with trait (identifying, for example, 15 markers out of 236 associated with yield (P<=0.01)) and separately using stepwise multiple regression to identify the significant predictors of trait (resulting in a residual set of 18-20 markers in the yield regression model). Multiple linear regression (MLR) was also applied using the marker set identified individually to be associated with trait. The authors report that the marker set found by stepwise regression gave higher r^2 following MLR than did a set based on the most highly individually correlated markers.

The authors attempted to control for multiple testing by reference to FDR and results identified to be of interest had P<=0.01 and FDR<0.2.

To have discovered 15 markers from 236 that are significantly associated at a nominal P value of 0.01 is noteworthy. That FDR is < 0.2 is unsurprising since 0.2 x 15 = 3 and we expect 236 x 0.01 =2.36 markers to have a P value of <=0.01. The question is whether it is credible that about 2% of a set of randomly selected markers genuinely associate with trait or whether the failure to correct for confounding factors has inflated the success rate? Although the authors demonstrate some agreement between the map positions of 8 of their significant markers and published QTLs I remain to be convinced that this experiment has really found QTL in such abundance. I would recommend a re-analysis with kinship as part of the model.

Rostocks *et al*. 2006 investigate the potential for association mapping in a population of 102 European barley varieties and a relatively high density marker array (500 SNP: Note that these SNPs are used later in this thesis on a different variety panel (Section 5.1 and Chapter 8); indeed this paper formed a proof of principle for that study). The decay of LD over

genetic distance is investigated as R^2 showing that LD extends no more than about 5cM in this panel. STRUCTURE and PCO analyses show the panel splits into winter and spring barley with the 2- and 6-row types less well distinguished.

To investigate the potential to detect Mendelian traits by association the authors attempt to locate un-mapped markers and demonstrate that about half of the markers could be mapped. The statistic used for mapping was R^2 without correction for structure.

An association mapping study using the mixed model was applied to the panel; unfortunately the detail of the origin of the kinship matrix is obscure but **Q** with *k*=2 was included as a covariate. The trait was 'seasonal growth habit' - which is the same thing as the winter/spring split – the experiment failed to find the known key genes for vernalisation (the physiological character underlying the winter/spring split) but did find strong association with winter hardiness; a preferred trait in winter barley also rarely found in spring barley.

The authors attribute their failure to find the vernalisation loci to epistatic interaction between two loci (*Vrn-H1* and *Vrn-H2*) which means some spring varieties have winter alleles at one or other locus. Additionally there are winter barley varieties which are facultative spring types (do not require vernalisation) which adds further error to the model. It could be suggested that the population division into two populations used in the MM analysis (since **Q** had *k*=2) matched almost perfectly any between sub-population polymorphisms close to the QTL; essentially population structure explained the trait variation. However, this is not correct here since the authors note the marker found to associate with winter/spring habit had a fixed polymorphism between the classes. The marker was ultimately found to be closely linked with a QTL for winter hardiness which is essential for winter sown barley, facultative or otherwise.

In a larger variety panel colleagues and I (Cockram, *et al.* 2008) have demonstrated the combined use of SA and GC in locating the two vernalisation loci. That we were successful where Rostocks *et al.* failed is likely to be because we incorporated SNP and other markers from within the candidate genes rather than relying on LD with adjacent genomic markers to detect the QTL. Despite this limitation the paper shows useful control of strong spurious associations since a reasonable number of genomic markers were included in the association panel.

Matthies *et al.* 2009 use a panel of 141 varieties to investigate associations between polymorphisms (12) of (1 →3),(1 → 4)-β-D-Glucan-4-glucanohydrolase and (1 → 4)-β-

Xylan-endohydrolase which are enzymes involved in cell wall degradation during malting which influence diastatic power[1], soluble nitrogen content[2] and total soluble sugar[3] in malt extract[4]. 24 genomic SSRs were used to define structure using 'STUCTURE' and the delta-K method of Evanno *et al.* 2005 is used to show that k=2, corresponding to winter and spring sub groups.

Association mapping is conducted using the **Q** matrix as a covariate in a GLM over the entire panel. In addition the authors partition the panel into winter and spring types and conduct association analysis within the sub-sets without attempting to account for structure in the model. The authors found association with some of the haplotypes of the genes and noted that the full panel with structure correction gave essentially the same result as the spring only panel without structure correction. The winter only panel showed no associations, possibly reflecting greater breeding effort with respect to the malting quality of spring barley.

Although the authors conclude this shows promise for fine mapping using association techniques it is similar to the Cockram *et al.* (2008) study in that the marker panel is likely to include SNPs close to the functional polymorphism and certainly within the relevant gene. The technique of partitioning the panel to avoid the confounding effect of structure is interesting but might erode power since a quantitative trait - such as those investigated here – would be sampled less often and hence the genetic effect would be less easy to separate from the environmental variation. It is possible that if the marker panel were less biased towards success the small number of SSRs used to define structure might have proved inadequate.

Haseneyer *et al.* 2010 use a panel of 224 diverse (geographical origin; Europe, Asia, N. Africa and America) spring varieties genotyped with 45 SSR markers (MAF =0.05) to investigate association with thousand-grain weight, crude protein content, starch content, plant height, and flowering time. The authors use the MM to simultaneously estimate variety effect on trait from trials (conducted at three sites in Germany in each of two years) data also

[1] Diastatic power is a measure of the enzymic activity of a aqueous extract from malted grain. Grain with high diastatic power can be used in combination with cheap starch (typically wheat or maize) in the production of grain whiskey or cheap lager. Diastase is a generic name for enzymes which convert starch to soluble sugar.
[2] Soluble nitrogen is a nutrient source for yeast fermentation.
[3] Soluble sugars are the raw material for alcohol production in fermentation.
[4] Malt is the name given to barley grains which have been partially germinated and then dried. The process solubilises protein, produces diastatic enzymes and begins to solubilise starch. Malt extract is a hot aqueous extract from milled malt.

incorporating a **Q** matrix for structure and marker genotype to examine association. The **Q** matrix from STRUCTURE is found using the delta-K method to have optimum *k*=2 which corresponds to a 6-row 2-row split in the variety panel. Association is found with all traits (P=0.05 – Bonferroni corrected).

Whilst the simultaneous modelling of trial, structure and marker effects is impressive the design of the experiment is suspect since much trait variation will be due to poor adaption of the variety to the German trial location. The authors allude to this in the context of extending the variation in the phenotypes and incorporating rare alleles in the study but the possibility that structural differentiation between varieties from different locations might spuriously contribute to association is overlooked. This is despite some evidence of structure within the two major groups identified. I suggest that 45 SSR markers may have been insufficient to model structure; additional markers both for association mapping and definition of structure/kinship would serve to answer the question. The authors note that QTL discovered in the study for plant height, yield components and grain quality correspond to published QTL which lends some support to the findings of the paper but it is not clear how likely this success would be if the associations had simply been distributed randomly.

Roy *et al.* 2010 report an association study comprising 318 wild barley accessions 558 Diversity Array Technology (DArT) and 2,878 SNP markers (MAF 0.1). The trait studied was spot blotch resistance.

TASSEL was used to implement a GLM incorporating a **Q** matrix from STRUCTURE with *k*=10. Significance was determined using an experiment-wise p-value threshold estimated by permutation. 13 QTL were found, 12 of which are mapped and two map within 3 cM of each other. Taking this as 11 QTL five are reported to be co-incident with published QTL.

The authors give the locations of previously published QTL (totalling 12) this allow one to ask the question ‘how likely is this success rate if the discoveries by association are randomly positioned?’ Using a Monte-Carlo simulation (10,000 permutations) and assuming the published QTL are on average 10cM long it can be shown that the observed result would have occurred on about 20% of occasions if the QTL were randomly placed. So the co-incidence with published markers lends some support to the claim that some of the discoveries reported are genuine.

Hamblin *et al.* 2010 report an association study in which 1816 barley lines from 10 U.S. breeding programs (different breeders) were genotyped with 1536 SNPs. PCA on a matrix of

Manhattan distances, which equates to simple allele sharing in this inbred crop, showed that the panel divided into 6-row spring, 6-row winter and 2-row spring. The volume of data overloaded STRUCTURE so *k* and **Q** were determined on sub-sets of markers randomly selected but separated by > 5cM. Forty replicates runs (burnin: 100,000 MCMC: 200,000) at each of *K* = 1-10. The appropriate value for *k* was determined using the delta K method. With an admixture model *k*=7 was considered best but assuming no admixture the best value was *k*=10. The 10 sub-populations corresponded to breeding pools the 2-row/6-row/spring/winter splits also corresponded to breeding pools. A cluster analysis was found to match well with the k=7 model.

LD decay was estimated in terms of average R^2 between 1cM bins and although there were differences between the sub-populations, decay was usually complete in range 20-30cM which is in contrast with Rostocks *et al.* 2006 (above) but more similar with the findings I report in section 5.1.2.2, Figure 5-3. The authors do not conduct an association study but conclude that there are good prospects for such experiments if population structure can be controlled either statistically or by partitioning of the data.

Sun *et al.* 2010 used 175 varieties originating from Tibet, USA, Canada, Australia, Europe and N. Africa genotyped with 42 SSR markers. The phenotypes were quantitative, comprising; leaf area , stem diameter, grains per plant, filled grains per plant, grain weight per plant, plant height, spikelets on main spike, grains on main spike, grain weight on main spike, length of main spike, density of main spike, length of the 1st internode, length of spike neck, and awn length evaluated at one site (three replications). The diverse nature of the panel and the modest field trial program are a concern since phenotypic variation may well be due to poor adaption to the growing conditions. Additionally the genetic component of phenotypic variation has been assessed at one site in one year; there are no details of blocking structure in the trial meaning the partitioning of variance components may be poor. Association has been measured using a GLM with structure implemented in TASSEL. No indication of the value of *k* is given but the authors indicated that population structure is mainly attributable to country of origin. Finally I am concerned that 43 SSR are insufficient to both define structure and conduct genome wide-association. The authors report multiple associations with each trait against a p-value threshold determined by permutation. Despite the observation that many of the discoveries are on the same chromosome or part of a chromosome as published QTL for these traits I am suspicious that the authors have used the

wrong alternate hypothesis in their conclusions. Validation of the discoveries in biparental mapping populations would be the one way to resolve these doubts.

4.1.2 Arabidopsis (self pollinating diploid)

In the model plant species, Arabidopsis, Aranzana *et al.* (2005) shows that population structure results in high levels of spurious association and that SA can, at least partially, control for this. 96 diverse lines were selected from sites across the northern hemisphere and were genotyped with 876 re-sequenced genome fragments which were scored as haplotypes and were in turn treated as alleles for the purpose of data analysis. Specific genotyping for specific polymorphisms around three pathogen resistance loci and a flowering time gene was also conducted. A previously known gene for flowering time was identified in the association study

The authors report that population structure was related to latitude; inclusion of samples from the extreme north of Europe led to spurious association at many loci (uncorrected, 7% of markers were significantly associated with flowering time at a nominal significance threshold of P=0.01). These false positives could be controlled either with SA or by excluding these samples. Effects of structure were also evident in association with one of the three pathogen resistance loci. The authors noted that although they detected a QTL that influences flowering strongly (~13% of variation accounted), loci of lesser effect would have been overlooked, this emphasises the need for larger test populations in future studies.

Zhao, *et al.* (2007) report a further development of the work of Aranzana *et al.* (2005). 100 additional resequenced fragments known to be in and around several flowering time related genes were added to the genotypes and the sample set was augmented with F2 individuals from mapping populations. This allowed both association mapping and a conventional linkage analysis to be conducted on the same marker set. Phenotypes examined were flowering time in response to day length, vernalisation and the gene expression of both *frigida* (a gene involved in vernalisation response) and *flowering locus c* a flowering repressor. In the association study structured association, correction using PCA and MM were compared. The MM used comprised kinship measured using both simple matching and excess allele sharing.

By analysing the fraction of phenotypic variation described by the various measures of population structure the authors are able to show that population structure describes more of

the phenotypic variation than the markers closely mapped to the trait. The effect of this is evident in the association study where high type I error is the main feature of uncorrected marker association. Application of statistical control increases power over GC and reduces the number of apparently significant loci.

Where comparison is made between the association mapping and the trait mapping studies, however, some of the closely mapped markers are found to be non-significant in the association study. The authors attribute this to the high proportion of phenotypic variation accounted by the population structure models. In essence they observe that it is sometimes the case that population structure and phenotypic state are perfectly associated so correcting for one leaves no residue of the other. This is consistent with the observation by Aranzana *et al.* in the same population that removing the extreme northern samples was as effective as SA in controlling for spurious association. The paper highlights the value of kinship as a co-factor for controlling spurious association. Interestingly greater power is achieved, for these phenotypes when kinship is estimated with simple allele sharing rather than excess allele sharing (or allelic correlation).

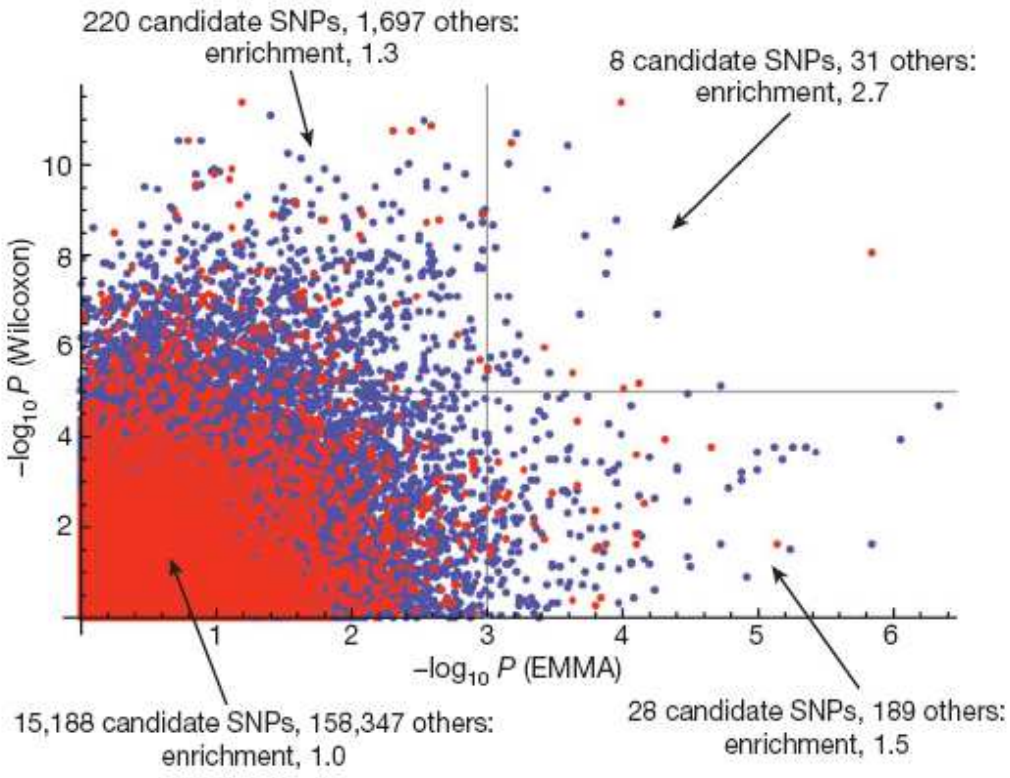
Recently an extremely comprehensive piece of research by Atwell *et al*. 2010 reported association mapping of 107 phenotypes, most with $h^2 > 0.8$ and all with $h^2 > 0.5$. The panel consisted of a core of 95 lines for all phenotypes with an additional 95 lines for some phenotypes. The marker panel consisted of 216,000 SNPs, which equates to 1 SNP per 500 bp in this small genome. This density is higher than is used in human studies, although the panel of genotypes is much smaller. Structure and kinship were controlled using MM. Of the 107 traits, association significant at a Bonferonni corrected p=0.05 were found for 33. For these discoveries the average number of significant markers was 2.7. This outcome contrasts strongly with many other studies where the number of discoveries as a proportion of total markers deployed is much higher. Although there are differences in LD decay profile between panels and species which will account for some of the difference it seems more credible to me that in this large, well resourced, experiment there has been good control of spurious association and the remaining power is insufficient to discover more than a small fraction of loci. This is probably (as the authors suggest) because the panel of lines is so small – the authors note about a two-fold increase in power for the traits for which phenotype is available for all 190 lines rather than the 95 core lines.

Compared to human studies the power seems high; the authors attribute this to the high heritability of the traits (environmental variation can be controlled in replicated trials) and the

genetic architecture of the traits discovered. In terms of genetic architecture the associations seem to be alleles of major effect (accounting for 20% or more of phenotypic variation), traits tested in the human context tend to be disease risk which is characterised by many genes each of small effect.

Atwell *et al.* go on to explore the question of true versus false discovery in a novel way because. for many of the traits candidate genes are known or proposed. The authors plot p-values from naïve association over corresponding p-values following the MM correction. Whilst there is only poor correlation between the two, the enrichment (that is the number of markers not close to a candidate gene over the proportion of markers close to a candidate gene) is greatest in the upper right quarter of the scatter, this suggests a screen which might be applied to assign priority to markers for validation.

The genotypes and phenotypes of up to 1000 lines will be added to this study in future; the increase in power will be very edifying.



.

Figure 4-1. Example of enrichment of true positives in the upper right of the scatter plot: associations proximate to a candidate locus are considered true and the proportion of such associations is highest in the upper right quarter of the plot. Red dots relate to markers within 20kb of a candidate gene; blue dots are other markers. The trait in this case is flowering time at 10C. (Reproduced from; Atwell *et al.* 2010).

4.1.3 Maize (naturally out pollinating diploid; often inbred by man).

The trait specific effect of population structure was also demonstrated by Yu *et al.* (2006) in a study of 277 inbred maize lines genotyped with 553 genic SNPs. These authors were demonstrating the use of MM for the control of population structure. They measured the effect of various control strategies on type I and type II error and showed that for flowering time and ear height, an improvement in power was obtained. No such improvement was observed in the third trait tested, ear diameter.

The MM approach allowed the use of kinship (in this case measured as co-variance between individuals) as a co-factor both alone and combined with a **Q** matrix from STRUCTURE (K=3; the authors report stable LnP(D) above K=2.). A very important outcome was that kinship alone gave greater improvement in statistical power than SA and almost as much

power as when structure and kinship were both included in MM together (see Figure 3-11, above). The authors suggest that, in inbred lines – and perhaps inbreeding species – kinship may be the more important cause of spurious association.

Another application of association mapping is presented by Weber, *et al.* (2007) who deployed 706 random SNPs to control for population structure and 48 markers in major maize domestication genes. The test species was teosinte (*Zea mays* spp *parviglumis*), the wild ancestor of the crop. The aim was to associate polymorphism in the domestication gene homologues with aspects of teosinte plant architecture and hence uncover the extent of the influence of these genes. This cannot be done directly in maize since these major genes tend to exist as single variants due to the bottleneck at domestication.

It is noted here (see also Remmington, *et al.* 2001) that LD decay in maize and teosinte is rapid, often complete within a gene. This contrasts to the situation in wheat where LD is reported to extend for as much as 10cM. The authors identify this as a distinct advantage in their study because associated markers tend to be physically close to causative genes, clearly realising this benefit is contingent on marker density.

The effect of population structure is evident in Q-Q plots showing inflation of test statistics. The authors simultaneously control for population structure, and kinship by applying MM using both a P matrix from principal component analysis as Price et *al.* 2006 (using the eigenstrat software from http://genepath.med.harvard.edu/~reich/Software.htm) and an estimate of kinship together with a spatial analysis of the field experiments used to measure the phenotypes. This is presented in a single model with candidate marker as a fixed effect. False discovery rate (<0.1) was used to deal with issues of multiple testing.

Ten significant associations with aspects of plant architecture were found, showing that what are viewed as critical traits for domestication may only be one of many possible wild phenotypes. The breadth of the effect could never be discovered in the domesticated species by these methods.

This is a very specific experiment directed at finding associations with polymorphisms within known genes and tested in an out-breeding species with very little LD and as such appears to have been very successful. The authors acknowledge the marker associations are ‘putative’, presumably meaning proposed, and need independent verification.

The observation that the bottle-neck of domestication has profound implications in association mapping has implications for how the technique is used. In the dissection of

genetic contribution to economically important phenotypes there may be merit in studying the domesticated species where these traits have been characterised in their range and in defined genetic backgrounds. For more fundamental understanding of gene interactions and the extent to which a single gene may influence many aspects of a plant's development the more polymorphic progenitor species might be better subjects for study.

Maize is both a grain and (in temperate regions) fodder crop. Andersen et al. (2007) report an association study attempting to fine map within a single gene, Phenylalanine Ammonia-Lyase (PAL). PAL catalyzes the first step in lignin biosynthesis in plants, lignin being the most indigestible component of plant tissue. The used sequence data from 32 lines which had been selected to represent a range of digestibility (in simple terms, % of dry matter digestible by a ruminant) and which comprised 19 flint and 13 dent types (a major population division in elite maize lines). Re-sequencing of the gene in these lines gave 39 SNPs defining 8 haplotypes. 101 genomic SSRs were used to define structure using STRUCTURE and 2 populations were found correlating with the flint/dent split. A generalised linear model (GLM) with no correction for structure was used to look for association of digestibility with the 39 SNPs and was compared with logistic regression in which **the Q** matrix was a covariate. The authors report an indel which associates with digestibility but sound a note of caution since this is present in only one line. There are multiple associations using GLM but these may be confounded with structure. The authors note that the GLM results may be correct but that the strong structural division in the data may mask this, a potential problem for association mapping in highly structured populations. I think that the panel size is too small to have achieved even the modest success claimed. The authors state that associations noted should be examined in larger, and more diverse panels.

Beló *et al.* 2008 report an association study of oleic acid content of oil in 553 inbred lines and 8590 markers. They find three QTL for the trait. One on Chrom 4 maps close to genes for fatty acid desaturation (*fad2*). Biparental crosses confirmed this and sequencing the gene from parental lines revealed a non-conservative amino acid change near the active site. The authors consider the possibility that differences in RNA expression for *fad2* might account for the change but find no evidence of this in a comparison between two contrasting backgrounds. This is considered to be the first report of the cloning of a causative polymorphism found by association in plants.

The approach to association mapping was unconventional in that the panel was divided into six sub-populations on the basis of STRUCTURE **Q** matrix and association was conducted without any correction for structure or kinship within each of the groups – thus six association analysis

were conducted, one per sub-population. One peak of association was found in three of the six sub-populations, and it was this which led to *fad2* and ultimately the causative polymorphism.

I would like to see the analysis repeated with a more conventional approach to the control of sub-population – perhaps the MM + K – if for no other reason than to check it works. Overall this is an impressive achievement which was probably made possible by a large panel size and a well defined trait.

4.1.4 Durum Wheat. Self-pollinating allotetraploid.

Durum wheat has attracted some attention for association mapping since (apart from being a locally important crop in its own right) it is a progenitor of bread wheat. It has a less complex genetic history thus having the potential to inform association mapping in the latter species.

As early as 2005 Maccaferri *et al.* reported on the extent of LD and the potential influence on measures of association due to population structure. The study utlised 134 varieties and 70 SSR markers. STRUCTURE found 8 -10 sub-populations. LD in the whole panel extended to 50cM with strong off-chromosome LD signals noted. When the data were partitioned into the identified sub-groups the extent of LD fell to 20cM and the off-chromosome LD was substantially reduced. The authors conclude there are good prospects for association mapping in the crop but that the effects of structure must be controlled.

Reimer *et al.* 2008 conducted an association study using 93 diverse (globally dispersed origins) durum accessions on the concentration of a yellow flour pigment. Phenotype was assessed in four separate field trials grown in Canada. Association with 244 SSR markers was assessed. Structure was defined using 28 unlinked markers (one per chromosome arm). The average max likelihood of the STRUCTURE model was observed at k= 5 but only 10,000 : 10,000 (burnin/MCMC) were used to optimise the model. Association was measured using GLM within the software package 'Trait Analysis by aSSociation, Evolution and Linkage' (TASSEL) Bradbury *et al.* 2007.. A permutation test gave an experiment-wise significance threshold.

The authors review QTL for the trait and include the published locations on a genetic map together with the locations of their marker panel. They note that 48% of significant associations are coincident with previously published QTL and propose this as confirmation that the association mapping has worked. However, closer examination of the figure reveals that 80 of the 244 markers fall into published QTL; using this as the expected distribution under the null, neither the distribution of all significant markers (31 fall into a QTL out of 81 observed) or

markers showing significance in all four trials (23 in QTL to 57 total) are significantly different from the expectation under the null (chi-sq 1 d.f. test). It is thus difficult to conclude otherwise than that either the association mapping has failed to bias the discoveries in favour of 'real' QTL or the original publications are not themselves mapping genuine QTL for this quantitative trait. I suspect the experiment has failed to favour existing QTL because of the extremely diverse origins of the varieties – this may have forced these essentially unadapted genotypes to exhibit new genetic mechanisms to influence the yellow pigmentation.

Maccaferri et al 2010 map yield components in a panel of 189 varieties of diverse origins. The varieties were grown in 15 field trials split (7/8) over two seasons. Each trial site was in a different Mediterranean country. The authors emphasise the different water stress experienced by the trials at these different locations and attempt to map drought tolerance as an additional factor. The marker panel consists of 186 SSR markers (MAF =0.1).

STRUCTURE was used to investigate sub-population structure and an unusual criterion was used to determine the appropriate value of *k*; the authors note that above *k*=5 the sub-population assignments of many individuals fall below 0.5; they view this as instability and adopt *k*=5 . This does not appear to be a particularly relevant criterion unless there is prior knowledge of sub-population structure.

Association mapping was conducted using TASSEL, the significance of marker-phenotype associations was tested using: (i) the general linear model (GLM) including the **Q** population structure coefficients as covariates, (ii) the MM including the **Q** population structure coefficients and the kinship matrices based on either Loiselle's kinship or (iii) that of Stich. [Loiselle *et al.* (1995); Stich *et al.* (2008)].

The phenotypes were averaged over trials partitioned into low medium and high yielding thus giving three 'sub-phenotypes'. Additionally the within trial averages were also used as phenotypes. Marker–trait associations were considered reliable when the significance was detected with the three association tests mentioned above. The following significance thresholds were considered P=0.05 and P=0.01 marker-wise level and P=0.05 experiment-wise level (Bonferroni corrected).

The authors use consistency over site and environment as a criterion for accepting a marker as a true association which may be a useful way to address the obvious criticism that growing un-adapted germplasm in diverse locations will distort the phenotypic variation. Indeed the

authors themselves note that as water stress increased the power of association mapping declines which they attribute to diverse coping strategies.

Of the 186 markers 24 are found to associate with one or more of the phenotypes, this seems a surprisingly high success rate, particularly as the individual markers generally account for less than 10% of variation. In fairness to the authors, they acknowledge that using a marker-wise significance threshold (not accounting for multiple testing) was a strategy for maximising discoveries. It would be better, surely, to use a lax q-value in these circumstances because they would have an estimate of false discovery rate amongst the discoveries?

Only two examples of correlation between the discoveries by association and previously published QTL are given. The authors conclude that with a better marker panel they would be able to do more. I wonder if they understate the importance of increasing the size of the variety panel?

Maccaferri *et al.* 2010a map resistance to 25 isolates of leaf rust (from both durum and bread wheat) in a panel of 164 varieties genotyped at 225 SSR loci (MAF >0.1). The resistance phenotype was measured in inoculated trials over multiple years and sites in both Italy and Mexico and in inoculated seedling experiments under glass. The approach to determining *k*, was as in Maccaferri *et al.* 2010, association analysis was also as in the previous paper except that experiment-wise significance following GLM was estimated by permutation.

The authors noted a previously reported resistance locus on chromosome 7B which was found in the variety *Creso.* Examining resistance associated with markers on 7B confirmed the presence of the resistance gene; by partitioning varieties according to haplotypes in a 20cM section of the chromosome it was possible to show that the most resistant varieties shared a three-marker haplotype with Creso – essentially fine mapping the QTL to this region. To separate this known resistance from other novel discoveries the main association analysis was conducted on varieties which did not contain the haplotype exhibited by *Creso* in the relevant portion of 7B (29 varieties); this was a three-marker haplotype spanning about 1cM.

Association mapping of the reduced panel showed some association consistent over isolates in seedling tests. However, the majority of associations were not significant against an experiment-wise threshold – the notable exception being in homologous regions of chromosomes 2A and 2B where association with resistance to one isolate from Durum and all

isolates from bread wheat was noted. The authors note that even the associations in linkage group 2 were not confirmed in field trials.

The conclusion is that association mapping allowed the confirmation and finer mapping of a resistance QTL on 7B and the discovery of a QTL in linkage group 2 which confers general resistance to leaf rust isolates originating in bread wheat.

The use of a permutation-based experiment-wise significance threshold which will have been less conservative than the Bonferroni threshold of the previous paper has allowed the authors to better differentiate between discoveries warranting further investigation and discoveries which may be little more than artefacts of multiple testing.

4.1.5 Potato. Out-pollinating autotetraploid.

D'hoop *et al*. 2008 conducted an association study of various morphological and cooking quality traits in a panel of 221 potato varieties genotyped with 250 AFLP markers, some of which were mapped. The potatoes were supplied by five breeding companies who also supplied phenotype scores so there is a lack of commonality in the estimate of phenotype. Decay of LD is estimated as R^2 using 221 on-chromosome comparisons. LD decay is complete at 8cM (P=0.01).

Association is measured by a GLM in which 'breeder' is a cofactor; no other adjustment for sub-structure was used. Significance has been declared at P=0.01 which the authors state includes a modest allowance for multiple testing, I suggest that the use of q-values would have been more objective.

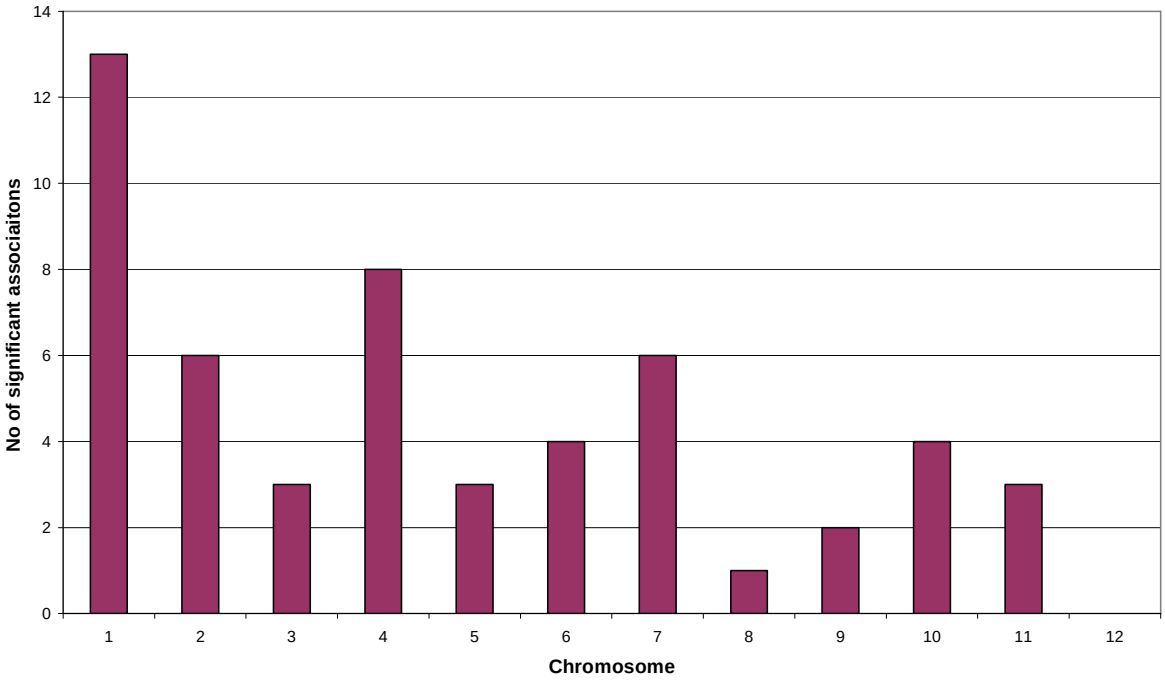


Figure 4-2. The distribution of statistically significant associations over the potato genome data. Souce; D'hoop *et al*. 2008.

The authors report 52 significant associations over all traits, they also list published QTL for the same traits and note there is coincidence between the two lists. However, taking an incidence of an associated marker and a published QTL on the same chromosome as a success and examining the distribution of successes under the null by MCMC (10000 permutations) for each trait I find only the specific gravity phenotype to show any evidence of enrichment ($P<0.001$). Otherwise the possibility that the observed numbers of matches occurred by chance cannot be readily dismissed. I also note the significant associations are not randomly distributed in the genome (Figure 4-2). This may suggest associations are being driven by some structural element in the population which may (for example) be related to loci on chromosome one.

Li *et al.* 2008 report association with fry colour – (a crisping characteristic), tuber yield and starch yield – of 243 individuals genotyped at 36 loci. The loci were predominantly sites of genes involved in sugar metabolism but included some non-genic markers. A total of 188 polymorphic fragments were found over these loci – some were polymorphisms within genes and some came from multi allelic markers. The MAF of polymorphic loci was restricted to ≥ 0.1.

The authors noted little evidence of structure - they found that the log-likelihood of the STRUCTURE model continued to rise over increasing *k*. I am not sure this is particularly strong evidence of absence of structure; it could actually be evidence of kinship. STRUCTURE might be finding related groups of individuals and interpreting them as population sub-structure. It would certainly seem likely that the multiple breeding pools would inject some structure into the panel and the authors acknowledge this by incorporating 'breeder' as a co-factor in a GLM. Significance was determined using q-value (0.05). Significant markers are subjected to stepwise multiple regression to identify multi-marker models better explaining phenotype. MM with kinship based on markers is used together with a MM with kinship and with breeder as a co-factor. The authors also conduct a naïve association for comparison.

Multiple associations were detected for fry colour and starch content but not tuber yield. The associations were found irrespective of regression model which is perhaps stronger evidence for the absence of structure; any general inflation of the test statistic may have been partially accounted by the use of q-values. Multiple regression found multi locus models explaining between 25 and 55% of variation.

Overall it is perhaps unsurprising that these markers, known to be close to genes involved in sugar metabolism, were associated with starch and fry-colour. Nonetheless, the authors report a successful procedure.

Li *et al.* 2010 continue from the previous study and look for epistaic interaction using a regression model with a between marker pair interaction term and breeder as a covariate. Q-value is again used for significance and FDR <0.2 is the threshold chosen. All possible marker pairs were considered. As a way to test for the potential effect of more subtle structural features in the panel the authors checked if the significant epistatic interactions persisted when the first 20 principal components of the relationship matrix were included in the regression model as co-variates. 50 epistatic pairs were identified, 45 of which persisted when the principal coordinates were added to the regression.

The authors added QTL to the regression model to test for confounding of epistatic and main effects and found a small proportion of interactions were not independent of main effects.

The paper gives a useful analytical framework for others to follow.

4.1.6 Wheat. Self-pollinating allohexaploid.

Ravel et al 2006 conduct an association study associating the protein content and gluten content of 113 wheat varieties grown in two trials (one per year for two years). Genotyping consisted of 24 genomic SSR markers used to define structure and three polymorphic markers known to be within *glu-B1-1* (coding for a component of wheat gluten) and *spa-B* (storage protein activator) which activates transcription of gluten genes. The objective of the paper was to determine which of three markers (2 for glu-B1 and one for spa-B would be the best predictor of gluten content.

The number of sub-populations was determined to be 5 on the basis of a peak of log likelihood over *k*. Individuals were then assigned unambiguous membership of the sub-population to which they showed maximal membership. The resulting population assignments were used as co-factors, together with year effects in a GLM to measure association with the three markers and a haplotype constructed from the two *glu-B1* markers. No association with *spa-B* is noted but protein content and gluten content both associate strongly with the *glu-B1* markers. The most significant association was with haplotype rather than individual marker alleles; haplotype accounts for 37% of variation in gluten content. The authors make a recommendation for the 'best' haplotype for high gluten content.

This is a robust piece of work and it is difficult to argue with the findings since the markers were already known to map to relevant genes.

Breseghello and Sorrels (2006) performed a limited association study of 62 SSR markers on chromosomes 2D, 5A and 5B with grain morphology and aspects of milling quality in 95 wheat varieties. The authors started with 149 varieties but removed 54 because they appeared closely related to others on the basis of genetic marker data, they refer to the process as normalisation.

Phenotypes were measured in field experiments at two geographical locations. STRUCTURE was used to determine cryptic population structure using 36 SSR markers distributed over all chromosomes except 3A and 6D. In order to determine the stability of the sub-population model the authors test randomly re-selected subsets of these markers. When they use fractional membership>0.5 as a criterion for assigning population membership they note that small numbers of markers are sufficient, as more confidence is required in assignment. As more markers are needed and 36 are insufficient to unambiguously assign all samples at fractional membership>0.9.

The authors use MM implemented in R to measure association with markers. Association was found between markers and aspects of grain morphology in a 10cM region of 5B and a 5cM region of 5A. Association with grain width was also noted over much of chromosome 2D.

The normalisation process (removal of genetically similar lines) seems attractive, in that a potential source of spurious association is also removed, i.e. it is a physical control for kinship effects. An alternative view is that there may be closely related individuals with differing phenotype which would be a source of important information on trait marker association. It would have been interesting to see the association mapping study conducted with and without these additional samples and the effect of statistical control for kinship could also have been investigated.

Crossa *et al.* 2007, report an association mapping study in which mapped DArT markers are associated with disease resistance genes in 170 breeding lines of spring wheat grown over the five years 1979, 1984, 1988, 1999 and 2004. No line is common to all years so MM is used to infer variety effects.

Population structure is measured using SRUCTURE on a year by year basis, presumably to look for structure within sets of lines within experiments. Two MM were deployed. In one model, BLUPs were biased by the fractional population membership of the line calculated by experiment and then associated with marker genotype. This was repeated as a separate analysis for each sub-population. In the second model the coefficients from the **Q** matrix were explicitly included in the MM used to calculate the estimates of genotypic effect on phenotype. Marker genotype was then associated with the predicted genetic effect.

Widespread association with disease state was observed, suggesting that population structure has not been controlled. When the locations of associated markers and previously reported QTLs were compared there appears to be some agreement and the authors cite this as evidence that the association mapping is finding genuine associations. However, it can be demonstrated by counting the number of previously published QTL listed by the authors and the number of associations they report for each trait that there is no evidence (Monte Carlo simulation; 10,000 permutations) to reject the null hypothesis that the reported matches have occurred by chance.

Stich *et al.* (2008) demonstrate the potential to apply MM in association mapping in wheat using both one-stage (association with phenotype is estimated in a single model including genotypic effects, components of error from the experimental procedure (rep, trial, block, etc)

and an estimate of structure) and two-stages (more conventionally; genotypic effect is estimated separately from experimental data and then association with markers is estimated in a MM including estimates of structure and/or kinship).

Kinship is estimated using pedigree data, estimates of IBD based on allele frequencies and also by the method originally proposed by Lynch (1988) and Melchinger *et al.* (1991) and described above (3.3.2).

The study was statistical in nature, attempting to infer the best approach to control type I and type II error from a very limited dataset. The authors conclude that MM combining both **Q** (structure) and **K** (kinship) matrices is appropriate in autogamous species. The authors recommend the use of the K_T kinship method based on REML estimate of P(two lines carry alleles IBS but not IBD) (3.3.2). This appears to improve power and reduce type I error. A two-stage approach to MM appears no worse than the one-stage alternative.

Apart from the K_T method for determining kinship this work essentially revisits the use of MM as a method for partitioning phenotypic variance to kinship, structure and then estimating the significance of marker genotype on the residual. It does serve to underline that there is a growing consensus that MM is likely to be the method of choice in inbreeding species.

Peng *et al.* 2009 report associations with Russian wheat aphid resistance in a panel consisting of 71, mainly Iranian, wheat varieties genotyped with 81 SSR markers. STRUCTURE is used to determine **Q.** TASSEL is used to implement a GLM with three thresholds for significance; marker wise, experiment wise by permutation and 'experiment-wise' it is not clear what this third threshold is (possibly Bonferonni corrected 0.05?). The author's do not report an association unless it is significant at all three.

Of 81 markers tested 29 are significantly associated with the infection symptom 'chlorosis' whilst 9 are associated with another symptom, leaf rolling. The associations are spread genome-wide. The author's conclude that they have found multiple new QTL. The experiment appears to lack the power to achieve this and I suspect that attempts to validate the QTL will be largely unsuccessful. A simple test for power – spiking the traits on the basis of marker genotypes – would have allowed the authors to at least measure how likely this level of success might be.

Yao et al (2009) conducted an association mapping study of yield components confined to chromosome 2A. 108 varieties were genotyped with 37 chromosome 2A SSR markers

together with an additional 85 SSR + 40 EST SSR markers which were used to determine structure. The objective was to fine map a previously identified QTL.

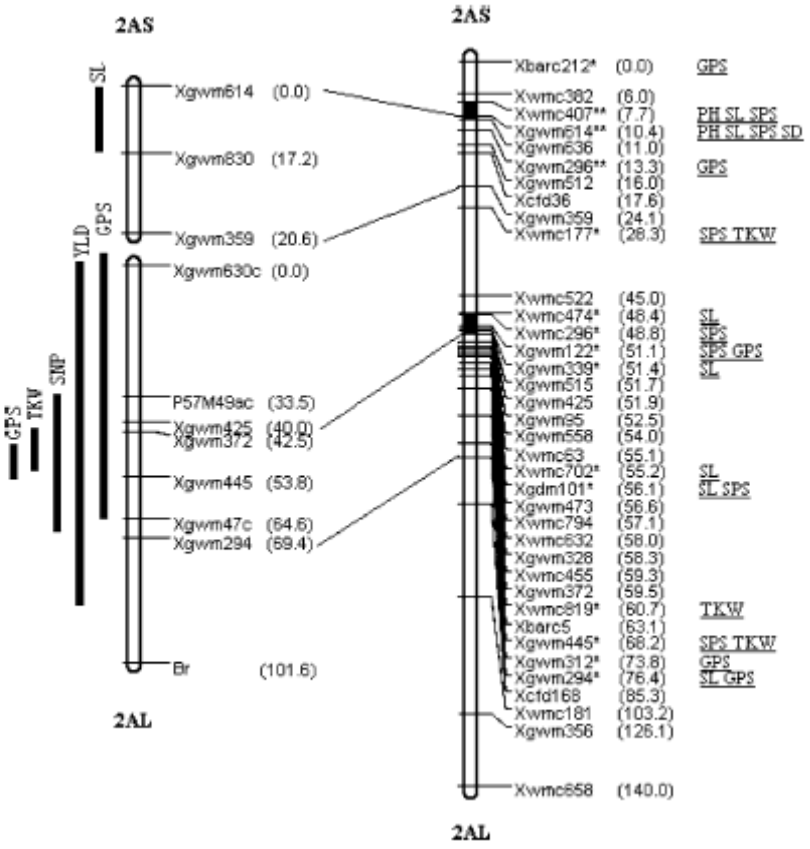


Figure 4-3. Distribution of markers used for association (right) and published QTL for yield components (left) in chromosome 2A of wheat. Significantly associated markers at P<0.05 and P<0.01 are indicated by * and ** respectively. Reproduced from Yao *et al.* (2009).

A MM was implemented in TASSEL using **Q** from STRUCTURE and K calculated using the internal TASSEL method. The number of sub-populations was determined to be *k*=9 on the basis that the improvement on log-likelihood for higher values of *k* was 'not significant'.

13 markers were found to be associated with yield using marker p-values. The authors assert that the significant markers map within existing QTL and into small genetic intervals. Thus an improvement in precision of mapping is claimed. However, Figure 4-3 shows the marker distribution; the improvement in precision may be due to the fact that the majority of the markers appear to map to the 10cM interval between 50 and 60cM..

Miedaner *et al.* 2010 report an association study of fusarium head blight resistance in wheat. A panel of 455 soft European wheats genotyped at 115 SSR loci is used. The authors use a one-step MM incorporating both trial design factors, kinship and marker genotype in a single

regression. A second model is fitted in which the interaction between marker pairs (epistasis) is also investigated. Experiment-wise p-value is Bonferonni corrected 0.05 for both models. I note that kinship is estimated using excess allele sharing relative to a maximum likelihood estimate of average allele sharing; negative kinships have been set to zero.

The authors find seven marker associations which match QTL positions in the literature and two potentially new QTL. In addition, two significant epistatic interactions are found each accounting for about 5% of genetic variance. The epistatic loci are not among the single marker associations. Although this appears to be a surprisingly high success rate the large panel size has enhanced the power. It will be interesting to see if the epistatic interactions are confirmed.

Neumann *et al.* (2010) map a variety of agronomic and morphological traits in 96 wheat varieties of geographically diverse origins (21 different countries). Phenotypes were measured in field trials repeated in up to 8 years (depending on trait) with 3 replicate plots per year all at one site in Serbia. This may have generated an abnormal environment effect on phenotype since the germplasm is un-adapted. DArT markers are used and the CYMITT map places 525 markers (MAF 0.05) into linkage groups. STRUCTURE was used to investigate sub-population structure and k=2 was determined using the delta K method; this seems unlikely given the diverse sources of the wheat varieties. I note that only 10000 cycles of both burnin and MCMC were used to determine the log likelihood estimates which casts some doubt on their validity. GLM with **Q** as a covariate and MM with **Q** and **K** (as estimated by TASSEL) were implemented within TASSEL to test for association. Each trait in each year was considered as a separate phenotype and a significant association was only recorded if a marker was associated at a marker-wise p-value of 0.05 over several (varies depending on the trait) trial years. The authors note that MM found about half as many associations as GLM. I think this may be partially accounted for by the poorly chosen Q matrix; the additional use of GC following GLM might have served to resolve this. 115 markers are found to be associated with one or more of the tested traits and the authors discuss occasions where their mapped locations coincide with published QTL. There appears to be no consideration of the null hypothesis – that associations were randomly distributed relative to the published QTL. Overall I am concerned that the lax significance thresholds (although the requirement for replication over years mitigates this to some extent) and the poor model for Q coupled with the distortion of phenotype due to poor adaption of the material make the findings suspect.

4.1.7 Plant association genetics conclusion.

Association mapping in plant species is developing rapidly but is still far behind comparable experiments in human genetics. The deficit is less in the number of studies as the scope of the studies themselves; panels of individuals and markers being generally two or more orders of magnitude smaller than those in human studies. Reported successes in association mapping in plants do generally lie under an alternate hypothesis but some studies present little convincing evidence that associations are due to LD with a genuine QTL. Even where association discoveries and published QTL are listed together there are several cases where the apparent confirmation of one by the other could easily be explained by sampling from two independent pools.

Nonetheless there is evidence that the plant science community has an advantage in terms of the accuracy with which the genetic contribution to complex traits can be measured, multiple observations over multiple sites for the same genotype is a luxury not available to human geneticists.

There is certainly evidence of structure in plant populations and a debate over the best method for control has developed in the literature. There is a growing preference for the use of mixed effect modelling with practitioners divided between users of TASSEL and a minority who use statistical software to construct their own models. There is also a division between two- and one-step modellers. One-step modellers being those who incorporate all the information from a trial programme (variety, block, site etc) together with kinship and marker genotypes in a single combined analysis and two-step modellers being those who determine a variety effect from trials data and then use that as a response variable in a simpler mixed model for association.

The marker density necessary to conduct a successful association mapping study in any species has not been considered in detail with most authors working with available data and concentrating on the measurement and control of structure.

Where LD has been examined in wheat it appears to extend to about 10cM (Table 4-1 summarises this). DArT markers (which I propose to use) have been used with some success by Crossa *et al* 2007 to confirm some previously reported QTL. In maize LD decay is more rapid which has led to association studies in and around known genes. As a prospect for

genome wide association mapping with limited markers, therefore, wheat appears the better choice.

Table 4-1. Published estimates of the extent of LD decay.

Range over which LD is observed.	Species	Reference
0-10cM	*T. aestivum*	Chao *et al.* (2007)
0-40cM	*T. aestivum*	Semagn *et al.* (2006)
0-5cM	*T. aestivum*	Breseghello *et al.* (2006)
0-5cM	*T. aestivum*	Somers *et al.* (2007)
0-0.5cM	*T. aestivum*	Tommasini *et al.* (2007)

Lacking in the current literature is a consideration of how best to exploit the advantages available to plant geneticists and the significance of the disadvantages. This consideration must include estimates of potential statistical power when panel sizes are varies, when effect sizes are varied and between quantitative and qualitative traits. Power, for example, is a routine prior consideration in human GWAS studies but is rarely considered by plant scientists. Panel sizes are limited only by budget in the human case but plant scientists may be constrained by the numbers of varieties available in seed banks and by the need to group together similarly adapted material.

This thesis attempts to cover these unaddressed questions by attempting to conduct association studies using historical phenotypes, seed-banked varieties an extensive simulation of statistical power.

5 Panels used for Association Mapping.

Three panels were available for this research; one consisting of barley varieties, one of wheat and one which was simulated under the Wright-Fisher neutral model and used to extrapolate beyond the limits of the other two panels. In this chapter the origins of the variety and marker panels are described, the marker panels are characterized and some analysis of population structure is conducted.

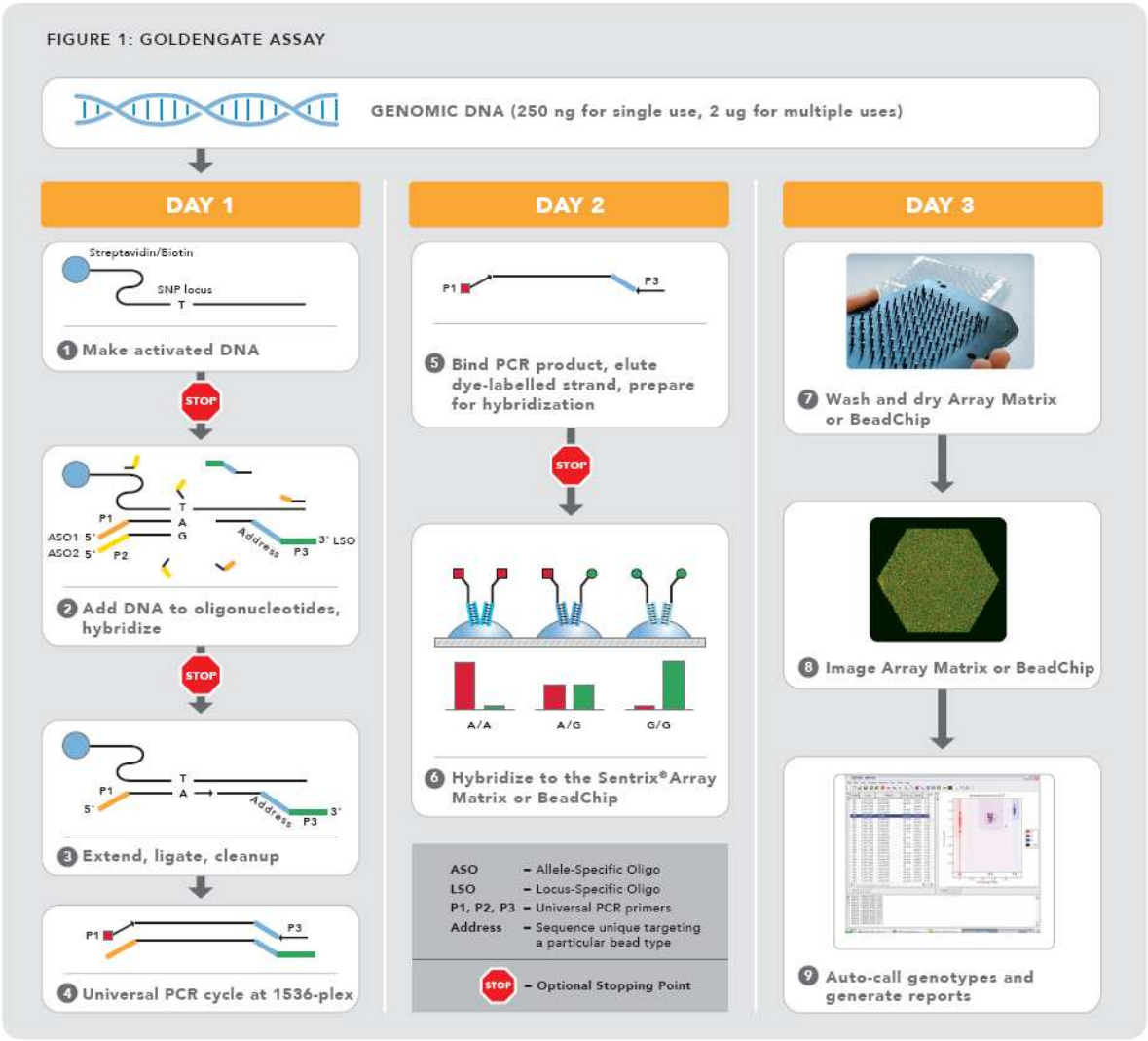
5.1 Barley Panel.

5.1.1 Germplasm and genotyping.

The barley dataset were those assembled by the AGOUEB consortium (http://www.agoueb.org/) for the specific purpose of conducting association mapping using a combination of historical phenotypes and some newly acquired phenotype data. These data were made available to me for association mapping of simple botanical traits.

The barley panel consisted of 490 UK barley cultivars from entries to national registration trials between 1993 and 2005 (Appendix 3). DNA was extracted from seedling leaf tissue using the Nucleoplex Automated DNA Isolation kit (Tepnel). A set of 1,536 EST-based SNPs were genotyped using GoldenGate BeadArray technology (Illumina) as summarised in Figure 5-1. SNP development is described by Rostoks *et al.* 2006. The variety set and the process of selection is described in more detail by Cockram *et al.* (2010).

The data had been pre-screened to have a minor allele frequency of 0.1 and the minimum data fill for markers was 96% and for individuals was 94%; this gave a panel of 1111 markers.



© *2006 illumina, Inc.*

Figure 5-1. The Goldengate assay relies on allele specific (ASO1/2) and locus specific (LSO) primers which together create a unique amplification product from each SNP allele at each locus. PCR is used to amplify these products in a multiplex (amplification of products from 1536 loci occurs together in one reaction). Dye-labelled DNA is used to differentiate between the alleles (step 4), the locus specific segment is then bound to a locus specific position on an array and the colour of dye at each array point allows the SNP genotype to be called (step 6). (Figure reproduced from: http://illumina.com/)

5.1.2 Genetic Map

The consensus map of this marker set described by Rostoks *et al.* (2006) was used as a basis for evaluating marker distribution, LD decay and subsequently for the location of candidate QTL following association mapping.

5.1.2.1 Marker distribution

The distribution of distances between adjacent marker expected under an hypothesis of random marker placement in a genome consisting of seven chromosomes with total map length identical to that used in this analysis[1] was compared with the observed distribution in the AGOUEB marker panel (Figure 5-2). The excess of both small and large distances suggests the markers are more clumped than would be the case if they were randomly distributed. This is consistent with both the origin of the SNPs which came largely from Expressed Sequence Tag (EST) sequence polymorphisms observed between nine barley varieties (Rostoks *et al.* 2006) and the nature of genetic maps. Since ESTs are cDNA (DNA complementary to mRNA strands), markers would be expected to be found in groups corresponding to genic segments of the genome. Microsatellite and other non-coding DNA would be poorly sampled and these unsampled regions probably contribute to the excess of larger distances between marker. Additionally, genetic distance is not linearly related to physical distance, particularly near the centromere, so where recombination rate is low markers spaced evenly on a physical map will appear to be clumped.

[1] The expected distribution was determined estimated by Monte Carlo simulation; 1111 markers were randomly placed on a genome matching that of barley in genetic length and chromosome number. The resulting between adjacent marker distances were recorded. This was repeated 1000 times to obtain the empirical estimate.

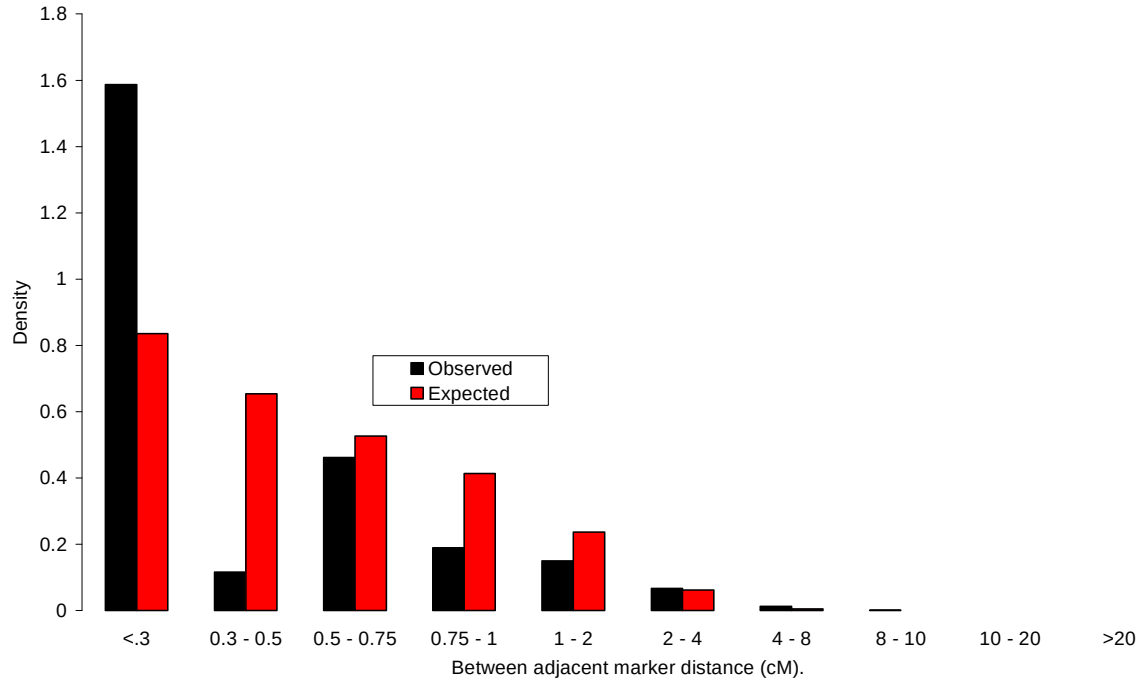


Figure 5-2. Density of between adjacent marker distances in the AGOUEB dataset compared with the distribution expected if the same number of markers were distributed randomly. The excess of lower and higher between marker distances can be explained in terms of; the marker discovery process which was biased towards genic regions resulting in clumps of markers closely spaced and an excess of large gaps between clumps and the non-linear relationship between cM and physical map positions. Note that the x-axis is asymmetrically scaled; column heights, however, are directly comparable.

5.1.2.2 Decay of LD and imputation of missing data.

LD, visualised as D' over between marker pair distance (Figure 5-3) shows LD decay extends to ~25cM in the combined winter and spring barley population. This distance encompasses almost all of the observed distances between adjacent markers which suggests that haplotypes observed in the genotypes could be used as a basis for imputing missing data provided the distances between adjacent marker were less than 25cM. Since few distances between marker exceeded 10cM (Figure 5-2) imputation of haplotype was used in preference to imputation based on, for example, *k* nearest neighbours.

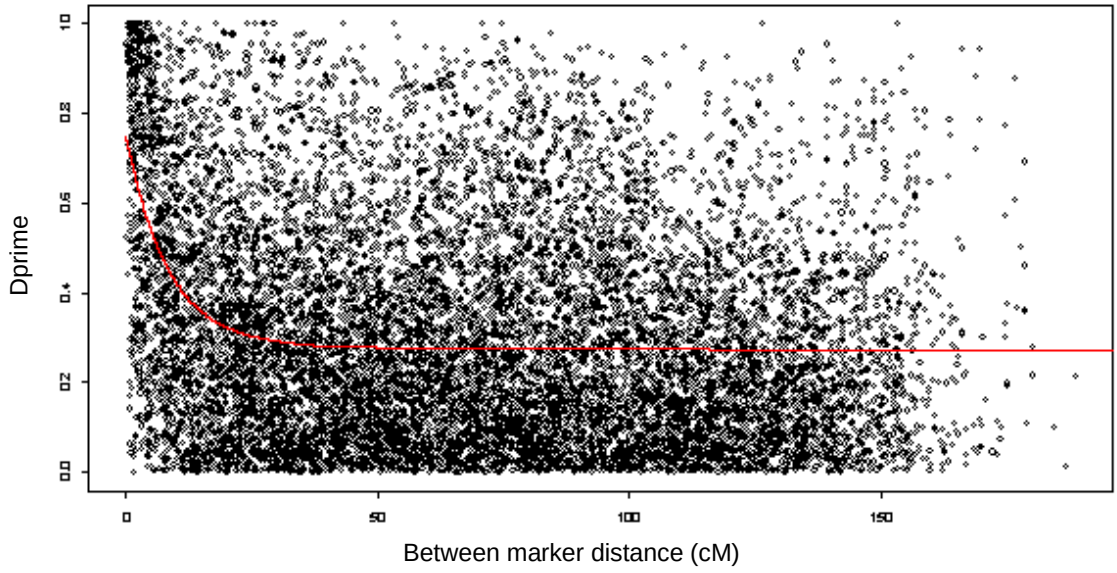


Figure 5-3. Barley population. LD between marker pairs measured as D' decays over genetic distance. The Malecot curve (red line) is asymptotic at about 25cM suggesting that an equilibrium state exists between pairs of markers at this separation (or greater).

Missing genotype data were imputed using the program SNPHAP (David Clayton: http://www-gene.cimr.cam.ac.uk/clayton/software/) which identifies haplotype frequencies in the data and then determines the most likely genotype for missing data-points. SNPHAP is written for out-breeding species and makes an assumption that the population is in Hardy-Weinberg Equilibrium (HWE). These data, being from fully in-bred varieties, are clearly not in HWE; this does not matter in this case because the excess of homozygotes simply removes uncertainty which would otherwise result from the inference of phase.

For imputation overlapping windows of 30 markers (20 marker overlap in either direction) were used for this process. The middle 10 data-points in each window were concatenated to give the final imputed data set. The first 10 and last 10 markers in each chromosome could not be taken from the centre of a window but instead the best estimate from the first and last window were used directly.

Each chromosome was imputed separately. No experimentally derived genotypes were substituted with imputed data; although subsequently I realised that the genotypes could have been quality assured by removing data points, imputing and checking that imputation and the

original genotype matched. In some circumstances it might even be preferable to replace a very unlikely genotype call with an imputed value.

5.1.3 Phenotype.

The historical phenotype data for the barley varieties comprised botanical characteristics used in the description of varieties for registration purposes. These characters are listed in 8.3 and are described in detail in the UPOV guidelines (TP 19/2; download from http://www.cpvo.eu.int/documents/TP/agricoles/TP_019-2rev_HORDEUM_VULGARE.pdf).

These botanical characters were originally chosen by taxonomists for their stability with respect to environment and the ease with which they can be assessed. The results are characters which have very high h^2. In addition yield data were provided.

5.1.4 Population substructure.

The recent genetic history of the barley population is summarised in 2.6.1 and the possibility that separate pools of winter, spring six-row and two-row barley will result in population structure has been noted. Four sub-populations might be expected in barley; winter/6-row, winter/2-row, spring/6-row and spring/2-row. And indeed, a principal component analysis of the genetic correlation between varieties shows 21.2% of the variation is accounted by the first two principal coordinates and graphically (Figure 5-4) four distinct clusters can be seen. AMOVA (Excoffier, 1992. Table 5-1) shows 34% of total genetic variation can be accounted using this population structure (F_{ST} =0.34).

That the population structure is not imposed by small groups of highly correlated markers can be demonstrated by excluding one marker from each pair of correlated markers (r^2>0.6). When this is done the same population structure is evident (Figure 5-4) and F_{ST} is not substantially reduced (Table 5-2).

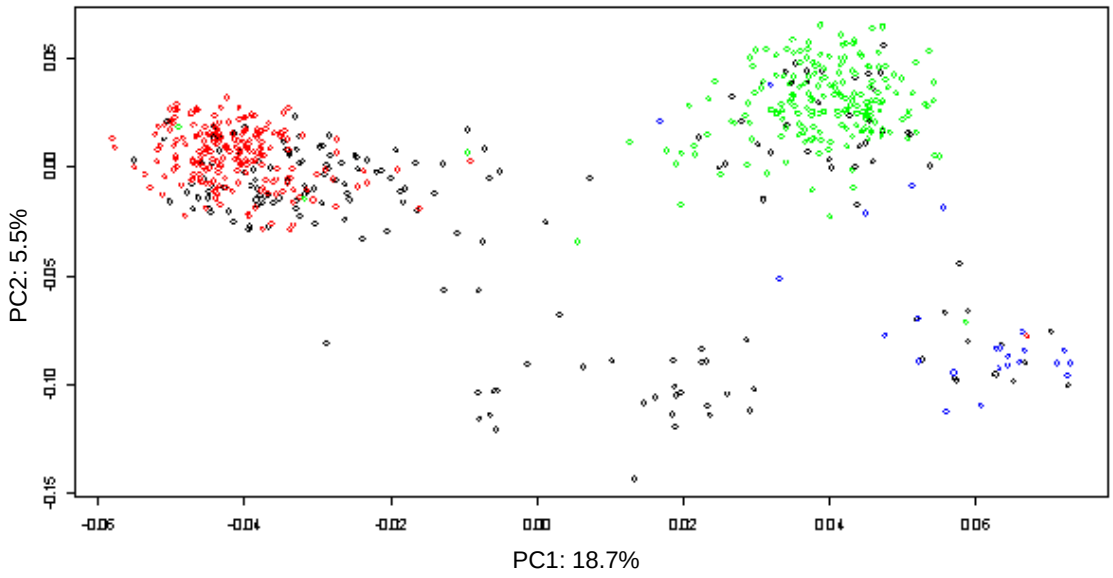


Figure 5-4. First two principal coordinates of a matrix of genetic correlation, calculated using only uncorrelated (r^2<0.6) markers, between individuals in the AGOUEB data set. Colours represent; winter/2-row : green; winter/6-row : blue; spring/2-row : red; phenotype not available : black.

Table 5-1 AMOVA of the AGOUEB dataset. Populations defined as winter/2-row; winter/6-row and spring/2-row all available markers used.

Source of variation	SSD	MSD	df
Among populations	19932.03	9966.02	2
Within pops	66234.76	158.46	418
Total	86166.79	205.16	420

Components of covariance	σ^2	%	p-value
σ_a^2	84.39	34.75	<0.001[a]
σ_c^2	158.46	65.24	
σ_T^2	242.85	100.0	

$$F_{ST} = \frac{\sigma_a^2}{\sigma_T^2} = 0.34$$

[a] H_1 ~ Observed covariance is greater than the covariance under the null (distribution under null estimated from 1000 permutations).

Table 5-2 AMOVA of the AGOUEB dataset. Populations defined as winter/2-row; winter/6-row and spring/2-row only un correlated (r^2<0.6) markers used.

Source of variation	SSD	MSD	df
Among populations	6041.4	3020.71	2
Within pops	27907.2	66.8	418
Total	33948.6	80.8	420

Components of covariance	σ^2	%	p-value
σ_a^2	25.4	27.6	<0.001[a]
σ_c^2	66.8	72.4	
σ_T^2	92.2	100.0	

$$F_{ST} = \frac{\sigma_a^2}{\sigma_T^2} = 0.28$$

5.2 Wheat Panel for Association Mapping.

As outlined in Chapter 1, the wheat panel and associated phenotypes were assembled from among historical variety collections and archived data. Thus the two potential limiting factors were availability of data and availability of seed.

5.2.1 Phenotype data.

Looking first at phenotype data; since the premise was that historical data were to be used and that the variety panel would come from historical varieties it was evident that the data matrix of historical phenotypes would be incomplete. At the very least a variety cannot have acquired phenotype data in years before it existed and in practice varieties have a short life in commerce (due to the process of supersedure by improved genotypes). The result was that in a database spanning thirty-one years and ~1200 wheat varieties the available data appears as a diagonal ribbon in a table of years by chronologically ordered varieties. The available data occupied ~3% of the potentially available data space (measured as total number of field trials multiplied by total number of varieties tested in the 31 year period). The consequence was that only a limited number of varieties ever appeared together in the same trial. Even the control varieties, grown in every trial, were changed every 5 to 10 years. The distribution of variety years in trial is illustrated in Figure 5-5.

An additional restriction came from mechanisms imposed to protect the intellectual property resident in the variety. Prior to registration on either the national list or the European common catalogue of varieties the genetic material is the property of the breeder. After registration the genetic material becomes publicly available but the right to sell seed is awarded to the breeder alone. Thus commercial benefit accrues to the breeder but other breeders may incorporate the new genetic material in their own breeding programs. The usual time for registration is after the second year of trial, thus varieties with three or more years in trial are publicly available and could be included in this study and named in results without restriction.

Taking the availability of data and the IP restriction together I chose to identify all varieties with more than two years in trial as candidates for inclusion in the study. This totalled 204 varieties.

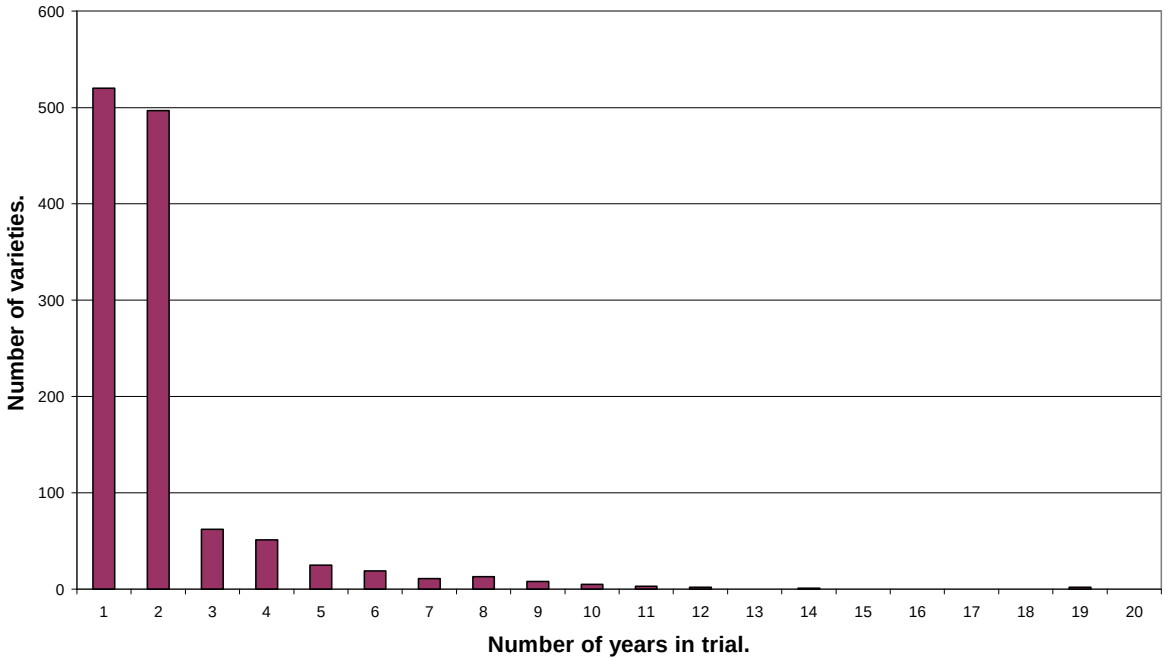


Figure 5-5. The majority of varieties are in trial for less than three years. Varieties which appear to offer commercial benefit after two years tend to complete four years in trial. Commercially successful varieties and controls continue in trial for longer still. The varieties Riband and Hereward were each present in 19 years of trials.

5.2.1.1 Estimating Variety Effects for Traits

The available data comprised computer records of variety means for yield from 1976 to 2007, each data point was the best linear unbiased predictor (BLUE) for the variety in a specific trial/year/fungicide treatment combination. From 1987 to 2007 similar computer records also exist for a range of other traits including protein content, Hagberg falling number and specific weight. These data are results of laboratory analysis of samples bulked from field replicates. Explicitly, a trial comprising three replicates of each variety would have yield recorded on each replicate allowing a BLUE for the variety in the trial to be estimated. An equal weight of grain from each replicate would then be combined to make a single bulk for laboratory analysis giving one estimation per trial.

Paper records from archive storage were recovered for the period 1976 to 1987 for protein content, Hagberg falling number and specific weight. Data from paper records were entered into the computer database and quality assured in one of two ways; where variety means for a year (i.e. mean over sites) had been recorded in the paper records the variety means for the

re-entered data were compared with the original data and discrepancies identified and reconciled by proof reading. Where no year means were available the trial data were re-entered in full a second time approximately two month after the first entry and discrepancies identified by direct comparison; these were then reconciled by proof reading. The volume of data is illustrated in **Table 5-3**. All data entry was carried out my me.

Table 5-3. The volume of historical trials data.

Trait	No Samples	No Years	No Varieties	No Trials	AVERAGES				
					Trial size (No. Varieties)	Trials per year	Replica-tions per variety[1]	Variety years in trial[2]	Variety replica-tions per year[3]
Yield	74332	32	1219	1472	50.5	46.0	61.0	2.2	22.0
HFN	21458	31	1324	586	36.6	18.9	16.2	2.0	8.1
Protein	17097	31	1145	691	24.7	22.3	14.9	1.8	8.4
Specific Weight	18452	31	1211	546	33.8	17.6	15.2	1.7	9.1

[1] This is the average number of estimations of the trait for a variety in the database, data from replication within a trial was not available.
[2] Many varieties are in only one year of trial.
[3] Not all varieties are present in all trials.

5.2.1.1.1 Trial Analysis

Mixed effects modelling was conducted using GENSTAT version 11 (Payne *et al.* 2009) used to estimate the BLUES of variety effect. The default (all factors except variety random) model was :

Trait = (constant) + year + trial +management + variety + variety x year + variety x management + management x year + year x variety x management + error.

Trial could not be used as an interaction term because it was recorded in a manner which made it unique to each year.

The magnitude of the dataset was such that the three way interaction sometimes had to be dropped from the models. Where model terms gave negative variance components these terms were excluded and the model re-analysed. The most complete model (**Table 5-4**)was then used to estimate BLUES for the variety effects for the trait; these were the values used in subsequent association analysis.

BLUES were chosen over BLUPS because in these data heritability **Table 5-5** is so high that the difference is negligible and the use of BLUES gives direct access to an estimate of significance of variety. This is the normal approach to variety trial analysis in the UK.

Table 5-4. Models used to estimate BLUES.

Trait	Random term model used to estimate BLUES. Variety was a fixed term.	p-value for significance of variety
Yield	Year + Management+ TrialID+ Variety.Management+ Year.Management + error	<0.001
Protein	Management + Year +TrialID + Management.Year + Variety.Management + Variety.Year + error	<0.001
SpWt	Management + Year + TrialID + Management.Year + Variety.Management + Variety.Year + error	<0.001
HFN	Management + TrialID + Year + Management.Year + Variety.Year + Variety.Management + error	<0.001

Heritability

Average trait heritability was estimated from the random model comprising only main effects, the denominator for the error term being estimated from the ratio of stratum variance to variance component for variety (as indicated in Appendix 1). Interaction terms were introduced to the limit of computing capability and heritability re-calculated, detail of the calculations are given below and in **Table 5-5**.

Main effects only (all random terms)

Source of variation	Expected MS
Year	$n_y\sigma_y^2 + \sigma_e^2$
Site	$n_s\sigma_s^2 + \sigma_e^2$
Variety	$n_v\sigma_v^2 + \sigma_e^2$
Management	$n_m\sigma_m^2 + \sigma_e^2$
Residual = V_e =	σ_e^2

Thus V_g = (VarietyMS-ResidualMS) / n_v
V_p = V_g + V_e / n_v
h^2 = V_g / V_p

In the conventional mixed model output the components of variation are already given, as discussed in Appendix 1 the difficulty is finding $n_{v,}$ etc in incomplete and unbalanced data. GENSTAT offers a solution since it provides both estimates of stratum variances and variance components allowing n_v to be recovered by simple arithmetic.

Table 5-5. Variance components and trait heritability.

	Variance components	Mean Square		(VarMS-ResMS)/V_g	V_g	ResMS/ n_v	V_g / (V_g +V_e)
Trait	Variety (V_g)	Variety	Residual	Average replicaton per variety (n_v)		V_e	Trait h^2 using all data.
Yield	0.3926	10.2650	0.4951	24.89	0.3926	0.0200	0.950
Protein	0.2255	2.0169	0.287	7.67	0.2255	0.037	0.858
HFN	1591	13287	1437	7.45	1591	192.89	0.892
SpWt	2.580	20.150	2.020	7.03	2.580	0.287	0.899

5.2.2 Sources of Seed.

Seed of historical varieties is usually stored by breeders although unsuccessful varieties may be discarded over time and mergers and acquisitions amongst breeding companies can complicate the process of finding which company has rights over which varieties. Breeders are under no obligation to provide seed samples. Additional to the breeders, the variety testing authority (DEFRA; Department of the Environment, Food and rural Affairs in the UK) also holds a reference sample of every variety tested. Seed of more popular varieties has been archived with seed banks such as the USDA small grains collection (http://www.ars.usda.gov/main/docs.htm?docid=2884), IPK Genebank (http://www.ipk-gatersleben.de/Internet/Forschung/Genbank) and John Innes Centre (http://www.jic.ac.uk/GERMPLAS/Index.htm).

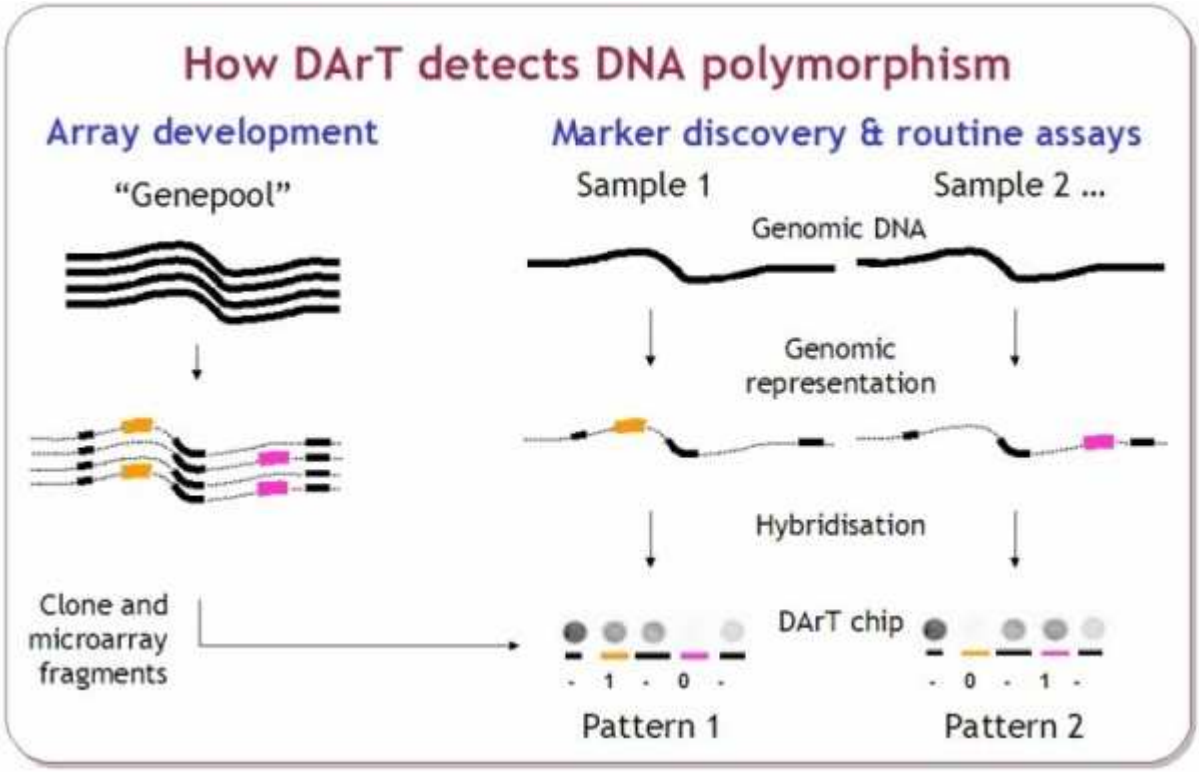
The route to obtaining seed samples preferred for this study was by application to seed banks who provided 5g of viable seed from their stock without charge; of the 204 varieties identified for the study I obtained seed of 119 in this way. The remaining 85 varieties were obtained by making accessions from the DEFRA reference collection, which is held at NIAB. However, this could only be done after I had obtained written permission for each named variety from the breeders concerned; a process which must have been irritating for busy breeders and was certainly administratively tedious for me.

5.2.3 DArT Genotyping.

Diversity array technology (DArT) technology is fully described by Jaccoud et al. (2001) and a detailed description is beyond the scope of this thesis, however a brief account is given below. In some respects DArT markers can be thought of as 'AFLP on a chip'.

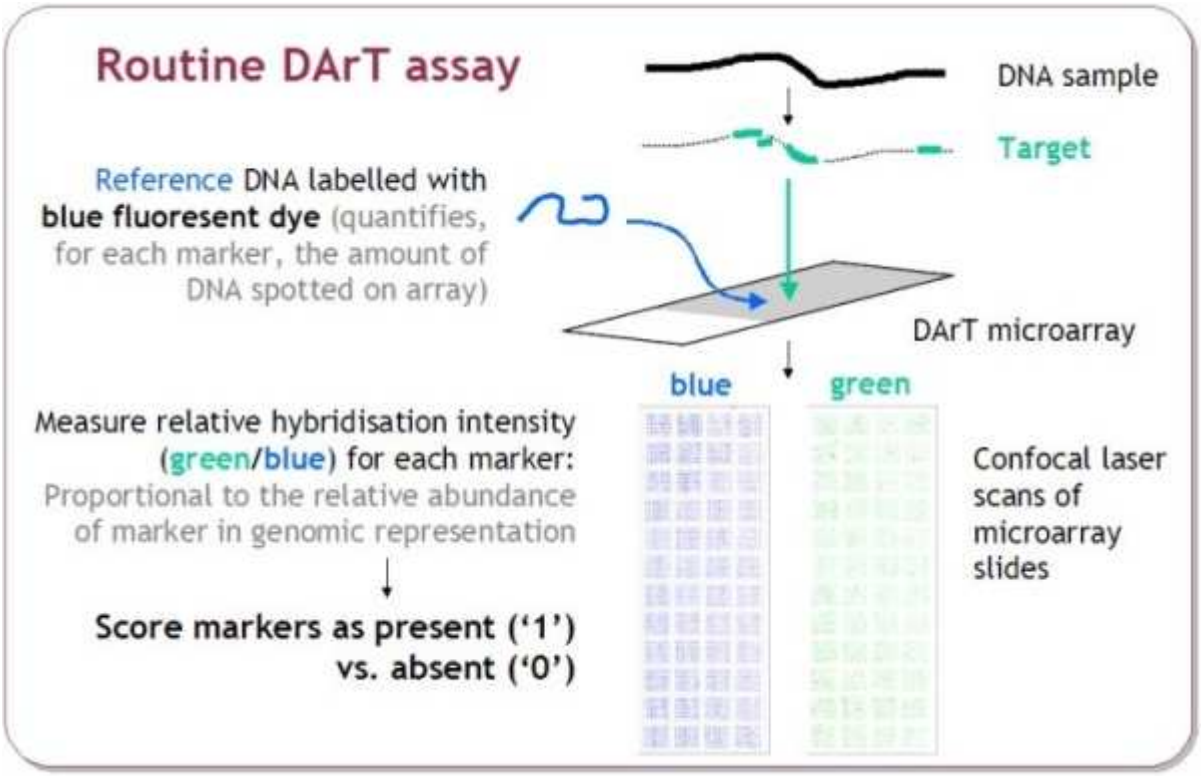
DArT samples the genomic DNA by fragmentation using restriction endonucleases. The majority of fragments obtained are universally present in the species under study but some polymorphisms are found due to SNPs in the endonuclease cleavage sites or gross Insertion/Deletion events in the DNA. The polymorphic fragments are identified by genotyping a 'discovery panel' of genetically diverse individuals. These polymorphic fragments are the basis of DArT markers, which are arrayed on a genotyping chip.

In routine analysis DNA from an organism is fragmented with endonucleases, labelled with a fluorescent green dye and then hybridised to the fragments on the chip. The chip is then flooded with fragments representing all the polymorphisms on the chip, these are labelled with a fluorescent blue dye. The chip is scored by recording the blue/green ratio at each point; high score for green indicates the polymorphism was present in the organism, high score for blue indicates that it was not. Figure 5-6 and Figure 5-7 give a diagrammatic representation of the process.



Copyright © 2009 Diversity Arrays Technology

Figure 5-6. Array development is from a wide sample of the species genepool. DNA is fragmented into monomorphic (black) and polymorphic (yellow and pink) fragments. Fragments are cloned and form the basis of a genotyping array. Polymorphic microarray fragments are identified by analysisng DNA fragments from individual organisms in a discovery panel, some microarray fragments prove to be monomorphic whereas others are present in some individuals but not in others, giving rise to different DArT genotypes. (Figure reproduced from: http://www.diversityarrays.com/molecularprincip.html).



Copyright © 2009 Diversity Arrays Technology

Figure 5-7. In routine analysis genomic DNA fragments from the test organism are labelled with green dye and compete with reference DNA (labelled blue) to hybridise with the microarray fragments on the array. The ratio of green/blue fluorescence for each marker is used to score the genotype. A reference is provided by the inclusion of some monomorphic markers for which the ratio of green to blue is assumed to be 1:1, significant deviation of observed ratio from the distribution in these reference markers is used to decide on the genotype score. (Figure reproduced from: http://www.diversityarrays.com/molecularprincip.html).

Akbari *et al.* (2006) describe the development of DArT for use in hexaploid wheat; this work was conducted on a biparental mapping population and a collection of 62 Australian wheat varieties which gave 2500 markers polymorphic in Australian and European wheat (referred to by Triticatre as Wheat PstI(TaqI) v2.3). In fact the DArT array used in the present study was an unpublished development from Wheat PstI(TaqI) v2.3 (described briefly at http://www.triticarte.com.au/content/wheat_diversity_analysis.html) incorporating an additional 2500 markers from other Triticum species and referred to as PstI(TaqI) v2.6. The DNA extracts were submitted for DArT genotyping using this, second, higher density panel.

5.2.3.1 DNA Extraction

DNA is most reliably extracted from green tissue so seeds were germinated prior to extraction. Germination took place in peat-based compost under glass-house conditions, twenty seeds of each variety were distributed in pairs in planting modules. Of the 204 varieties sown 35 failed to germinate.

Two phases of DNA extraction took place, the first in April 2008, was conducted by Richard Horsnell, a technician at NIAB, who intended to use the resulting genotypes for assessing genetic diversity of a large collection of European wheat cultivars with the intention of establishing a core collection of wheat varieties capturing a large proportion of genetic diversity. As outlined below the first phase produced DNA which could only be partially genotyped due to both variable concentration and mechanical shearing.

I carried out a second extraction in June/July 2008 to obtain the final DNA used in the study.

Although the duplicate extraction and genotyping delayed the research by several months it did provide two laboratory replicates of the extraction and genotyping process which could be used to check for evidence of misidentification of samples.

DNA extraction I

Approximately 10mm of the first true leaf of a single plant of each variety was taken into a 1.5ml Qiagen sample tube held in a 8x12 matrix. Each tube contained an acid washed tungsten carbide ball. Three samples from each leaf were taken to ensure material was available for additional extractions if required. Samples were then frozen at -80C prior to extraction using the Qiagen DNeasy extraction system (http://www.qiagen.com/products/default.aspx Cat No. 69181) an in accordance with the manufacturer's instructions.

DNA Quality Control (Phase I)

The DArT genotyping platform requires relatively high concentrations of DNA (>50 ng/µl) and the extract must be free of contaminants which would inhibit the restriction endonuclease MSE1 which is used to fragment the DNA prior to DArT genotyping.

Previous work (White *et al.* 2008) in our laboratory had successfully used the Qiagen system for DNA extraction from wheat leaf tissue as a precursor to DArT genotyping so QC was restricted to checking the DNA concentration using a spectrophotometric analysis of the light absorbance by the sample extract at 260nm (NanoDrop 2000 system: www.nanodrop.com), result summarised in Figure 5-8. Following analysis samples were diluted to have DNA concentration in the range 50-100 ng/µl. Seven samples with DNA concentration in the range 0-20 ng/µl were sent for genotyping without re-extraction.

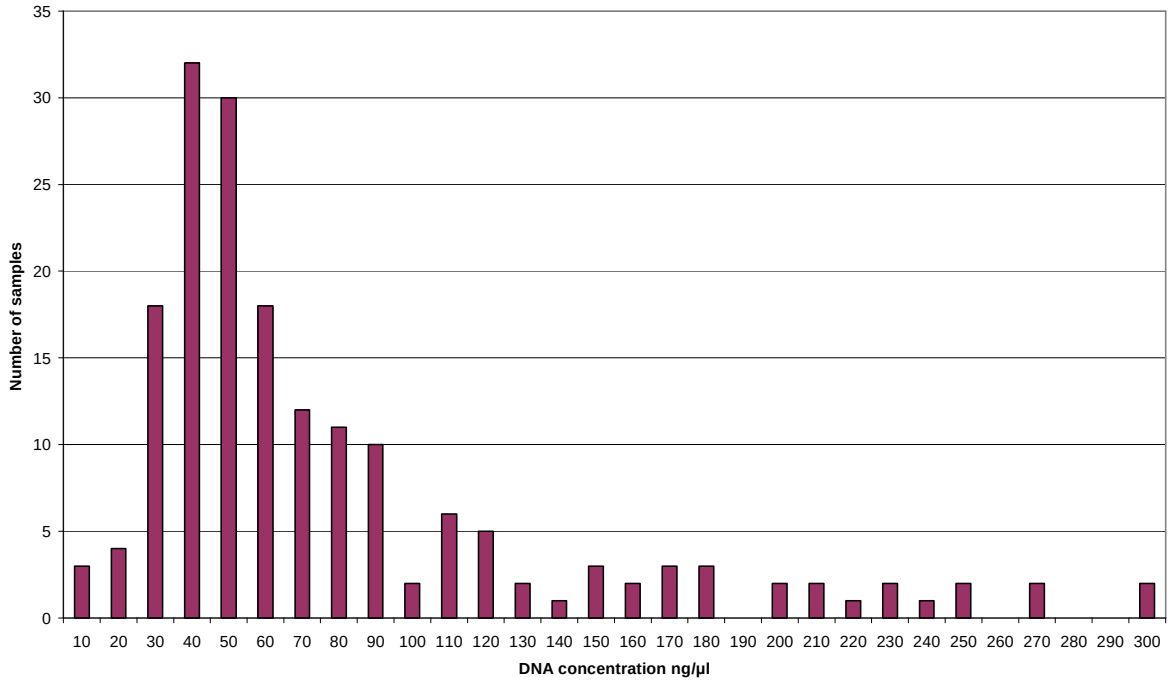


Figure 5-8. *The distribution of DNA concentration in samples from the first phase of DNA extraction. Samples were diluted as necessary prior to dispatch for DArT genotyping.*

DArT Genotypes obtained from Phase I

The DArT panel used (PstI(TaqI) v2.6) contains 5000 markers from *T. astivum* and other *Triticum* species. The genotypes received from the first phase of analysis contained data from only 504 markers. The supplier does not report genotypes of markers which are monomorphic in any panel but even allowing for this the paucity of data was of some concern, particularly as genotypes at 520 polymorphic marker loci had been obtained in a smaller set of UK wheat varieties in the earlier study (White *et al.* 2008) genotyped on the . Wheat PstI(TaqI) v2.3 platform which contains only 2500 markers. Enquiries with the

supplier revealed that their internal QA of the DNA supplied had suggested the DNA was of variable concentration and partially degraded (see Figure 5-9). They agreed to reanalyse a re-submission of DNA without further charge.

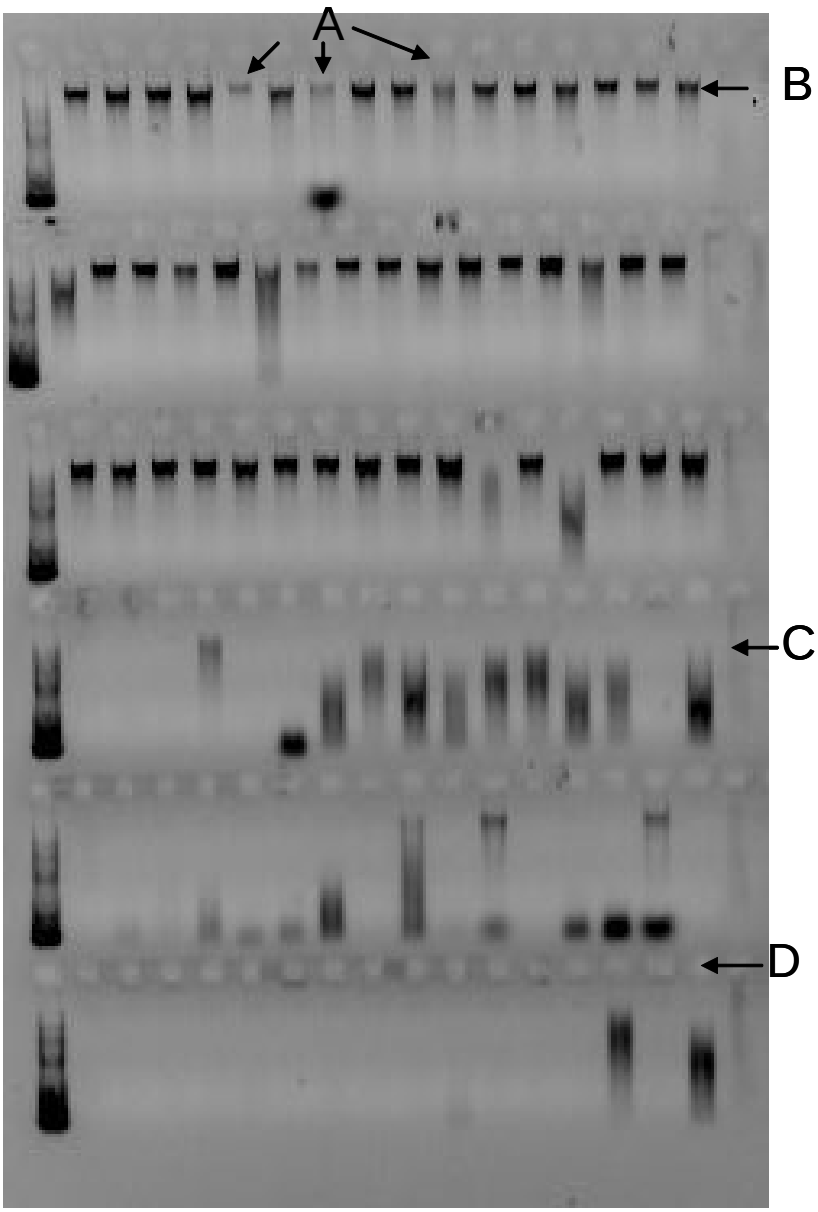


Figure 5-9. Example gel image provided by the genotyping sub-contractor showing; Variable DNA concentration evidenced as variable band intensity (e.g. three bands labelled A are of lower intensity than the others in the row.) Degredation of sample evidenced by the contrast between the first row where at point B all samples exhibit a single clear band of genomic DNA, and the fourth row (C) where no band of genomic DNA is evident but where DNA has separated into a 'smear' suggesting the sample comprises randomly fragmented DNA.

In the figure, twelve groups of eight samples have been examined, two in each of six rows. Samples were injected into the agarose at the points where the pale rectangles can be seen (example indicated by D). Samples migrated down the gel by electrophoresis and DNA in the sample was separated by molecular mass, the leftmost sample in each row is a DNA size standard.

Conclusion from Phase I extraction

It was evident that our QC of the DNA had been inadequate; the NanoDrop method indicated we had ample DNA whereas the contractor asserted there was insufficient in many samples for analysis. Equally serious, we had not expected to have problems with DNA shearing but there was good evidence that many samples had been degraded in this way. I concluded I should re-extract DNA from fresh plant material and perform QC checks similar to those used by the sub-contractor. This would give me full control of the process and give me data I could exchange with the sub-contractor in the event of further difficulties. Following discussion with the sub-contractor I decided to use a different DNA extraction procedure.

5.2.3.2 DNA extraction (Phase II)

Approximately 10mm of the first true leaf of a single plant of each variety was taken into a 1.5ml Qiagen sample tube held in a 8x12 matrix. Each tube contained an acid washed tungsten carbide ball. Three samples from each leaf were taken to ensure material was available for additional extractions if required. Samples were then frozen at -80C prior to extraction using the modified Tanskley method given in Appendix 5.

Attempts were made to extract DNA from seed material for varieties which had failed to germinate. Both the Tanskley method and the Qiagen DNeasy system (following manufacturers instructions) were used but insufficient DNA was recovered from this source to allow DArT genotyping to proceed.

Where re-extraction exhausted the reserve leaf material I obtained new leaf material by germinating seeds in petri-dishes on damp filter paper.

DNA Quality Control II

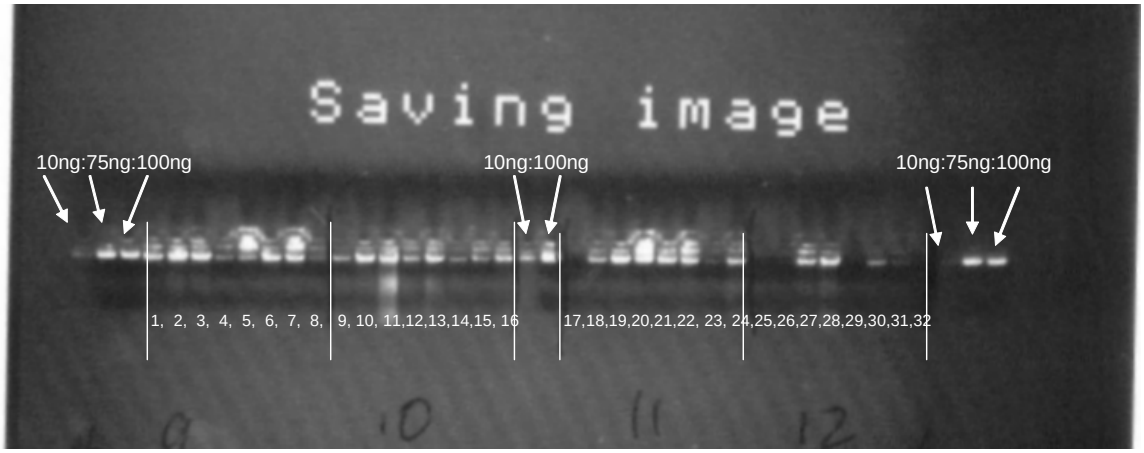


Figure 5-10. *Example of QC of DNA extracts. In this experiment the concentration of DNA is being crudely estimated by comparing the fluorescence from the DNA in a known volume of extract to the fluorescence from a known concentration of viral DNA, following agarose electrophoresis. The minimum standard is 50ng/µl. In this example extracts at positions 4,5,8,9,12,14,15, 17,23,25,26,29,30,31 and 32 were judged to contain insufficient DNA and were subsequently re-extracted. The bands marked 10, 75 and 100ng relate to viral DNA corresponding to extract concentrations of 10, 75 and 100ng/µl.*

To ensure extracted DNA met the previously stated requirements of concentration and freedom from enzyme inhibitors, an aliquot of each extract was visualised in UV light with ethidium bromide after electrophoresis on 1.5% agarose in 0.5xTBE buffer (pH 8.0). The intensity of fluorescence of the genomic DNA band was compared with injections of Phage Lambda DNA of concentrations 10, 75 and 100 ng/µl. (See example in Figure 5-10). Where the concentration of extracts was found to be too low a repeat extraction was made from the reserve sample of plant material.

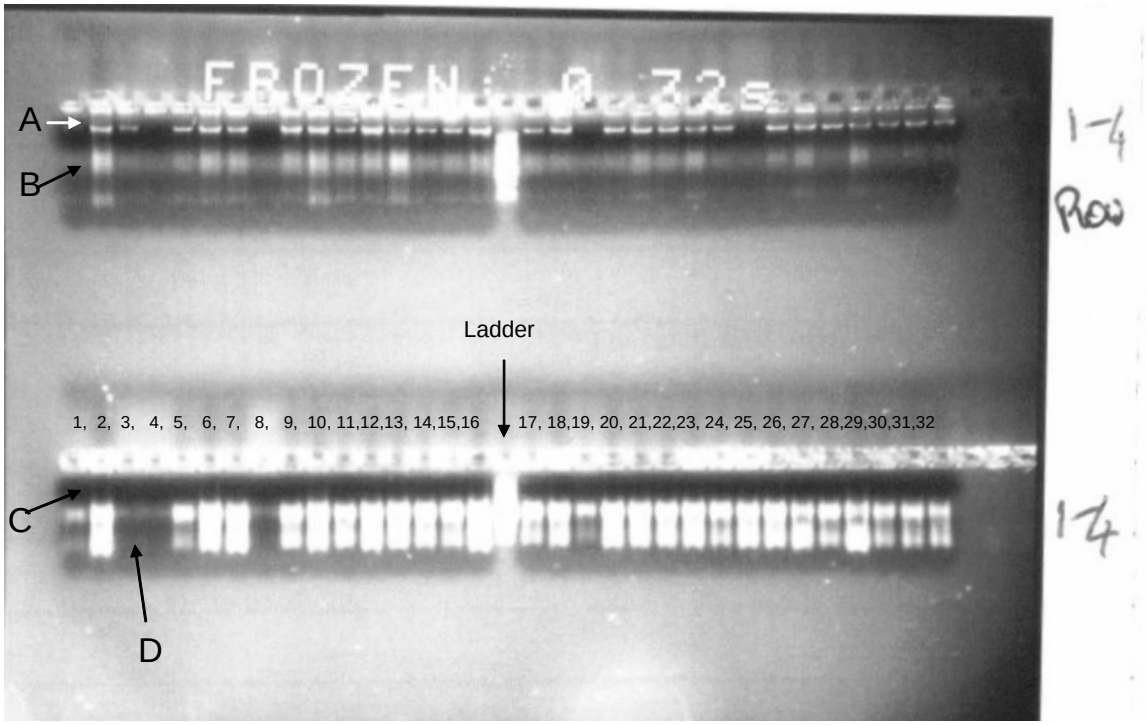


Figure 5-11. The upper row consists of undigested DNA extracts and the lower row consists of matching aliquots after MSE1 digestion. The DNA has migrated down the gel during electrophoresis; smaller fragments of DNA will have migrated furthest. The critical points to observe are that in the upper row a sample exhibits a bright clear band of high molecular mass genomic DNA (A) and that there is little evidence of smaller DNA fragments which may be the result of mechanical damage during extraction (B). The ladder contains a mixture of DNA fragments in the range 500kb to 10,000kb, it serves to demonstrate the fragmentation of the genomic DNA during digestion. Note that the genomic DNA band is less mobile than the largest ladder fragment but that following digestion the genomic DNA band has disappeared (C) and has been replaced by a wide band consisting of DNA fragments of less than 10kb; this indicates successful digestion. (D) indicates two samples (3 and 4) where there had been a failure to extract sufficient DNA. Samples 25 is anomalous in that there appears to be no genomic DNA but the digestion has been successful, this is attributable to the small volumes being handled, there has been a complete failure to transfer sample to the injection mixture in the upper row. The presence of adequate DNA concentration in this sample was separately confirmed by reference to the corresponding QC experiment (example in Figure 5-10).

Two cycles of re-extraction took place until at least one sample from each variety had sufficient DNA concentration for DArT analysis.

To ensure the DNA was of sufficient purity for MSE1 digestion an aliquot of each extract was digested with the enzyme and the product compared, following electrophoresis on agarose as described, with an undigested aliquot. A 1 kilobase DNA ladder was used as a common reference, example results are given in Figure 5-11.

Although replicate extraction was required to obtain DNA from all varieties no extract failed to digest with MSE1.

The second phase DNA extracts were posted to the sub-contractor in August 2008.

DArT Genotyping (Phase II).

DArT genotypes were received on 20th November 2008. The samples had been genotyped with 1294 markers out of 5000 in the marker panel, the remaining markers being monomorphic or having failed the sub-contractors internal QA system. This was consistent with previous experience (White *et al.*, 2008) in which we obtained data from 520 markers in a panel of 2500.

Summary of Genotype Data (Phase II).

Genotypes at 1294 markers were reported. Minor allele frequencies for each marker are summarised in Figure 5-12. Proportions of missing data for each marker and for each variety are summarised in Figure 5-13 and Figure 5-14 respectively. 1044 markers had a minor allele frequency ≥ 0.1 and were included in the study. I applied a minimum standard of 80% data-fill, no markers were excluded against this standard, however the variety Tanker had only 71% fill and was excluded.

Figure 5-15 shows the distribution of genetic distances between marker pairs observed and expected under an hypothesis of random placement[1]. There is evidence that the markers occur in clusters since there is an excess of low between marker distances and an excess of large between cluster distances. Akbari *et al.* (2006) observe that *'it seems that DArT markers have a stronger tendency than SSR and AFLP markers in particular, to map to gene-rich telomeric regions'* and this clustering is also reported in the development of a DArT panel in rye (Bolibok-Brągoszewska *et al*., 2009). In 2006 Wenzel *et al.* reported clustering of markers in a marker map of the barley genome incorporating DArT, SSR and RFLP markers. In this case some clustering was attributed to the disparity in physical and genetic distance in the centromere and also to a bias in the DArT markers towards the distal regions of some chromosomes.

[1] The expected distribution was determined estimated by Monte Carlo simulation; 1111 markers were randomly placed on a genome matching that of barley in genetic length and chromosome number. The resulting between adjacent marker distances were recorded. This was repeated 1000 times to obtain the empirical estimate.

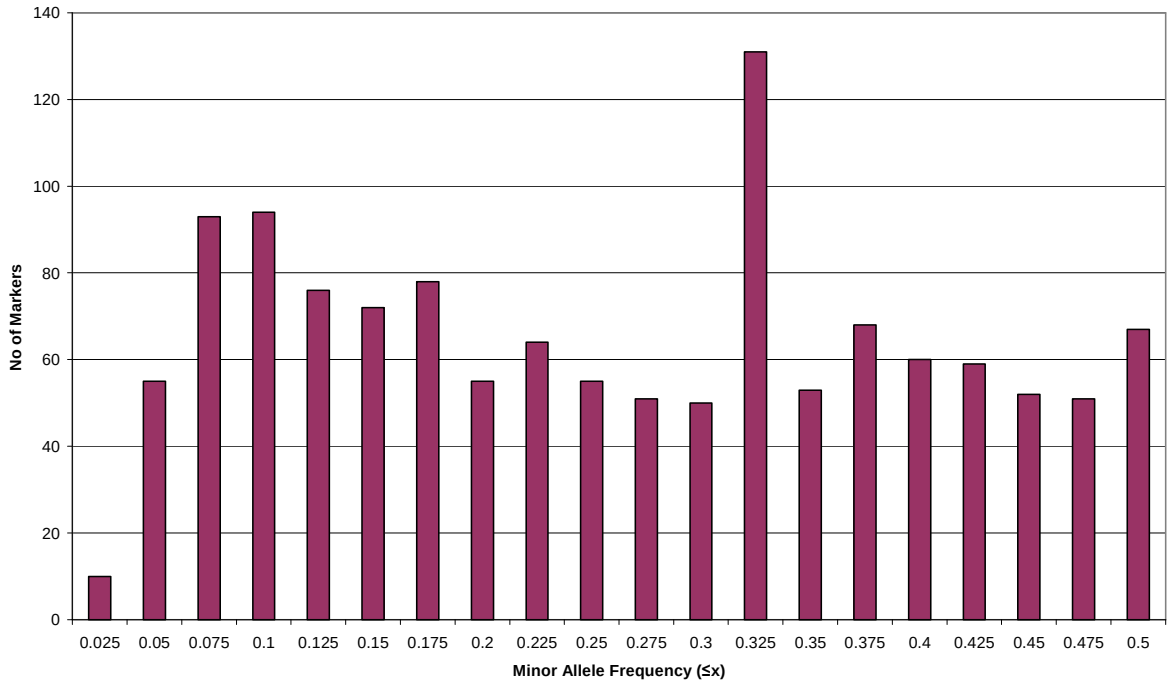


Figure 5-12. Distribution of minor allele frequencies in the genotype data received from phase II DNA extraction.

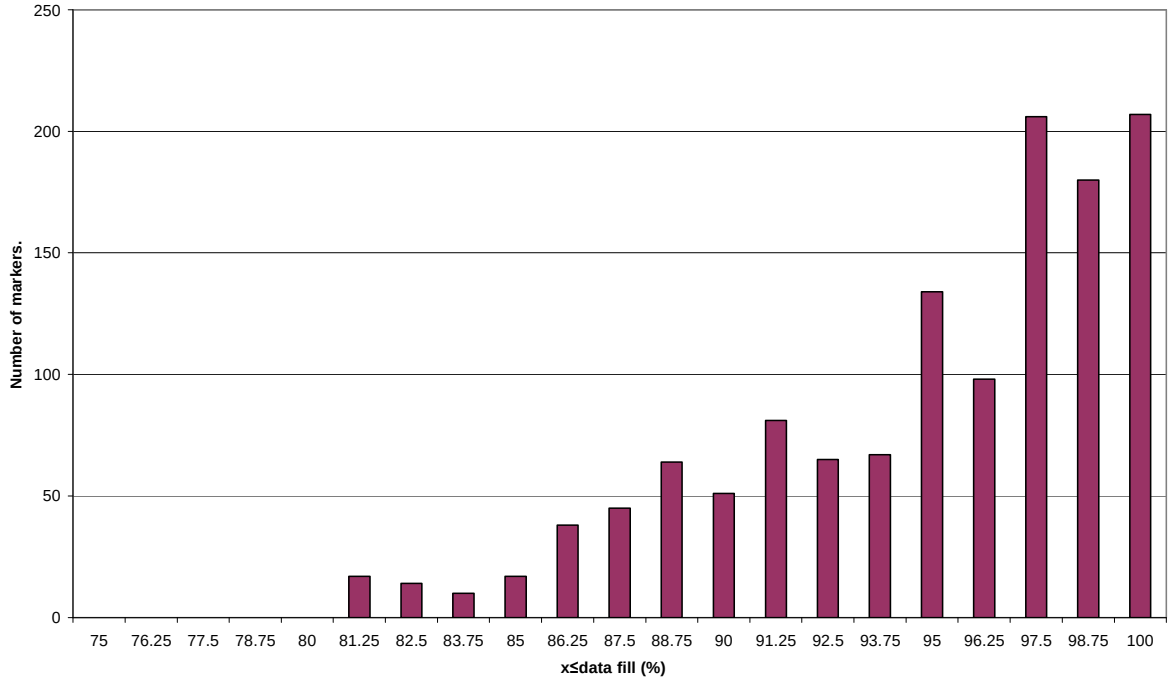


Figure 5-13. Data fill by marker for the genotypes obtained following phase II DNA extraction.

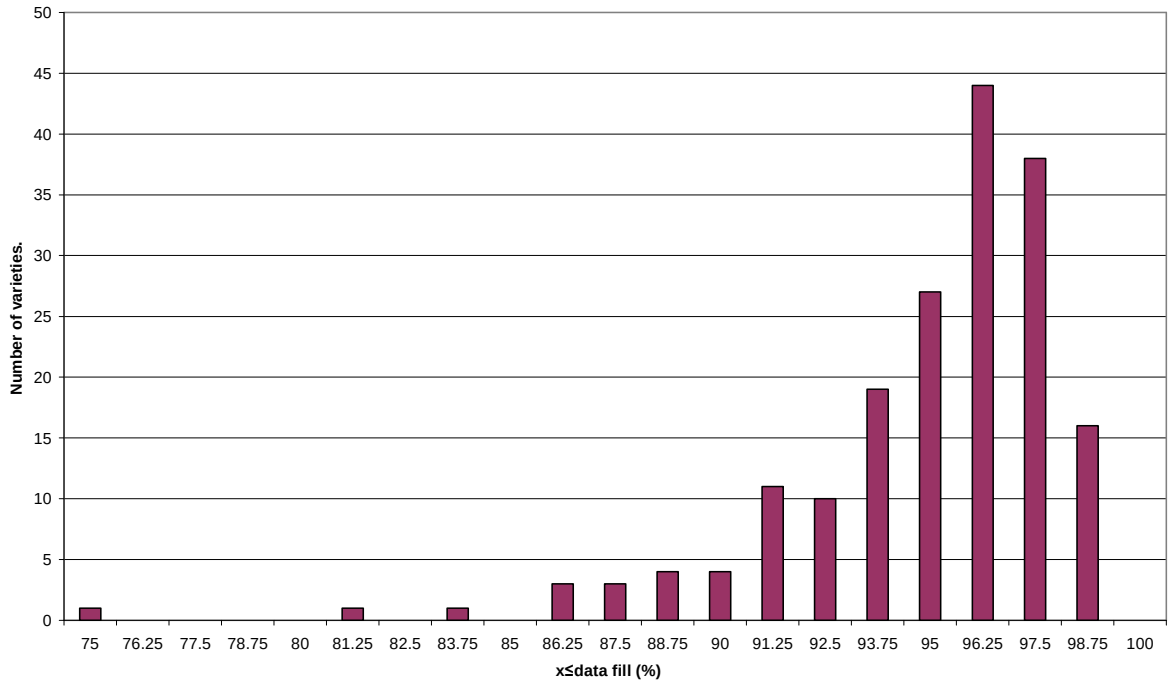


Figure 5-14. Data fill by marker for the varieties obtained following phase II DNA extraction. The variety Tanker had only 71% data fill and was eliminated from subsequent analyses .

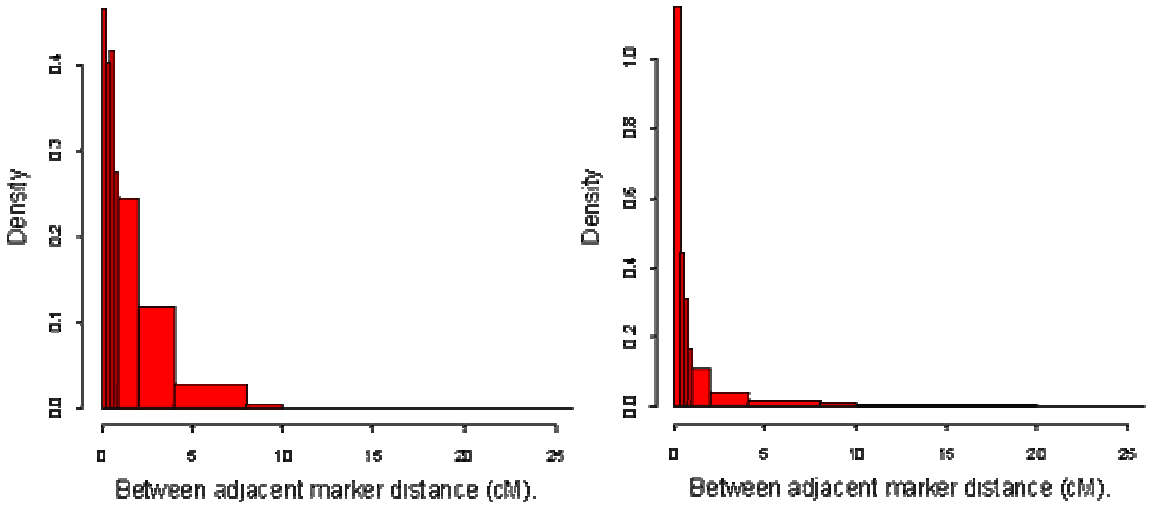


Figure 5-15. Density distribution of between adjacent marker distances in the wheat dataset (right) compared with the distribution expected if the same number of markers were distributed randomly (left). The excess of short inter-marker distances and the slight excess of large inter-marker distances may be attributed to marker location being biased toward genic regions as a consequence of the way markers are discovered.

5.2.3.3 Genotyping Error

In human genetics it is common practice to screen for genotyping error by testing for HWE at each marker locus (Hosking *et al*., 2004; Xu *et al.*, 2002). Xu *et al.* also advocate: Single blinding such that the genotypes are called without knowledge of case/control status; inclusion of blanks at random positions amongst the samples for genotyping; duplicate samples; independent scoring of alleles where possible and avoiding lack of balance with respect to missing data between cases and controls, re-analysing where necessary.

Testing for HWE was meaningless in these data since the varieties are fully inbred. It was, however, possible to incorporate two of the suggestions made by Xu *et al:*

Single blinding. Genotypes were analysed by a third party with no prior expectation of the outcome.

Duplicate samples. Partial over-lap of varieties over two studies and the (retrospectively) fortuitous re-extration of DNA gave some replication within the study. The samples sent for genotyping following the second phase of DNA extraction were divided between two 96 well plates. Two varieties (Beaufort and Bouquet) were common to both plates, they were labelled with variety name in the first plate but with a code in the second. Comparing genetic distance (1-simple allele sharing) between the pairs of genotypes gave zero mis-matches and one mis-match in the 1294 available markers respectively.

Considering occasions where duplicate analysis of varieties had taken place due to re-extraction and resubmission or in the previous (White, *et al.* 2008) study, **Figure 5-16** shows that the vast majority of duplicate analyses resulted in very small pairwise genetic distances between duplicate samples. However, a number of supposed duplicates did not match; Table 5-6 shows, for each such variety, the pair-wise distances between the available genotypes; it is reassuring that the final extraction run is almost always one of a matching pair, allowing the results for the final run to be accepted as the genotypes for use in association analysis.

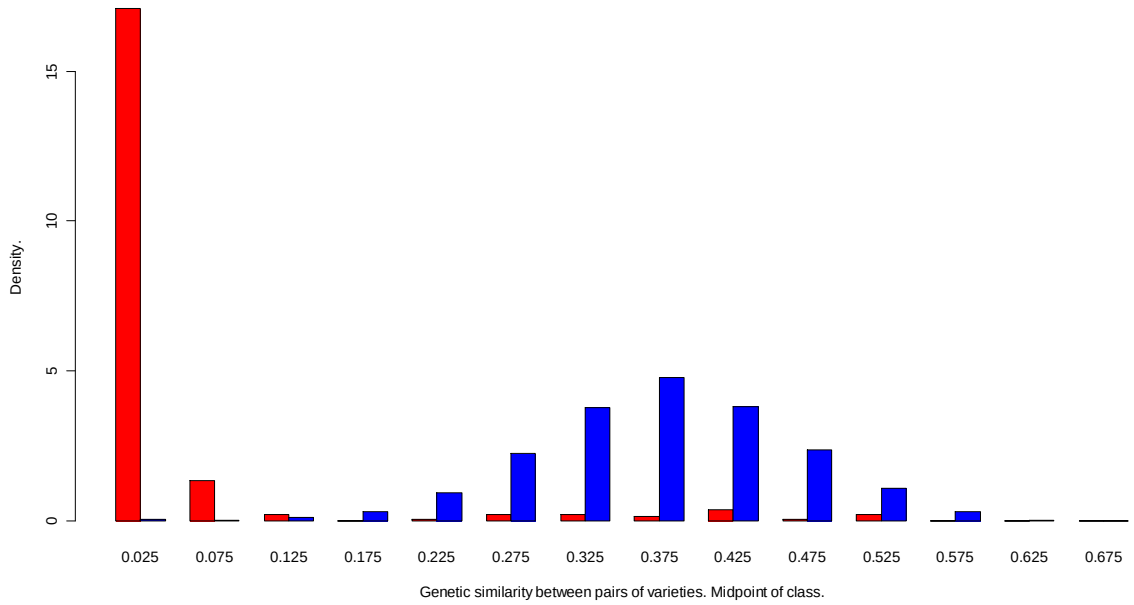


Figure 5-16. The pairwise genetic distance between varieties. Red bars are comparisons between replicate analyses of the same variety, DNA from different sources/extraction runs. Blue bars are all pairwise comparisons. Generally the agreement between replicates of the same variety is very good (distance <0.05). In most cases where replicates do not agree, and where three or more replicates were obtained, I was able to find two replicates where the pairwise difference was <0.05 (see Table 5-6).

Table 5-6. Genetic distances between replicate genotypes of varieties estimated as 1-(simple matching coefficient) for which at least one replicate pair had a difference greater then 0.05. Original diversity study(A), Phase I DNA extraction (B) Phase II DNA extraction (C). The source of the sample used in the final association study has been underlined. (*) indicates varieties for which the genotype could not be confirmed by replication, in these cases the genotype from Phase II DNA extraction was used since it had proved most reliable overall.

1-(simple matching coefficient)	Variety	Source; Rep 1	Source; Rep2
0	Admiral	A	C
0.3	Admiral	B	C
0.39	Admiral	A	B
0.43	Apollo	A	B
0.34	Apollo	B	C
0	Apollo	C	A
0.09	Armada	B	C
0.54	Ashby*	A	C
0	Atou	C	B
0.32	Atou	A	C
0.15	Avalon	A	B
0.09	Avalon	A	C
0.01	Avalon	B	C
0.11	Brigand	A	B
0.09	Brigand	B	C
0.01	Brigand	A	C
0.06	Battalion	A	C
0.06	Battalion	B	C
0	Battalion	B	A
0.06	Cantata	B	C
0.06	Cantata	B	C
0.14	Chatsworth*	B	C
0.1	Claire	A	B
0.06	Claire	A	C
0	Claire	B	C
0.01	Fenman	B	C
0.06	Fenman	A	C
0.08	Fenman	B	A
0.09	Flinor	A	B
0.03	Flinor	B	C
0.06	Galahad	A	C
0.08	Galahad	A	B
0	Galahad	B	C
0.29	Hyperion*	B	C
0.32	Mega	A	B
0.01	Mega	B	C
0.38	Mega	A	C
0	Mercia	B	C
0.08	Mercia	A	B

0.08	Mercia	A	C
0.1	Rapier	B	C
0.06	Rapier	A	B
0.01	Rapier	A	C
0.22	Spark	A	B
0.29	Spark	B	C
0.01	Spark	A	C
0.43	Steadfast*	B	C
0.51	Steadfast*	A	B
0.43	Steadfast*	A	C
0.06	Villein*	B	C

5.2.3.4 Imputation of missing Genotypes

Figure 5-22 shows the decay of D' over genetic distance. In this panel LD decay over the favoured map reaches its asymptote at between 5 and 10cM, however Figure 5-15 shows that ~ 25% of between adjacent marker distances are greater than this suggesting that imputation based on haplotype may not be the best approach in this case. For this reason missing genotype data was imputed using a genome wide estimate of similarity (Manhattan distance) and the k-nearest neighbour algorithm within the software 'TASSEL' (from: http://www.maizegenetics.net/index.php?option=com_content&task=view&id=89&Itemid=119) this implements the method of Cover and Hart (1967). In this case 5 nearest neighbours were found and imputation was on the basis of un-weighted average. The imputed genotypes took the form of fractional allele calls which therefore reflected the certainty with which the imputation had been made. Rows (varieties) with greater than 20% missing data were removed from the dataset at this point (see; Figure 5-14).

5.2.3.5 Summary of Wheat Panel.

As a result of genotyping failure, and non viable seed of the 204 varieties initially identified as avaialble for study; 167 were included in the final panel (see Appendix 4). The marker panel consisted of 1044 markers of which 358 were mapped to a unique chromosome position using the MetaQTL consensus map (5.2.4 below). Table 5-7 summarises the process.

Table 5-7. Summary of the development of marker and variety panels.

Variety Panel		Marker Panel	
Varieties:	Number	Markers:	Number
In trial since 1975	~1200	On genotyping chip	5000
In trial>3years	204	Rejected by sub-contractor[1]	3706
Obtained from seedbanks	119	MAF <0.1	250
Obtained from DEFRA	85	Available (mapped+ unmapped)	1044
Failed to germinate	35	Mapped	358
Removed due to poor data-fill	1		
Available for association mapping	167		

5.2.4 Genetic Map

No consensus map existed for the DArT panel at the beginning of this research, however the genotyping subcontractor supplied alignments from nine mapping populations (http://www.triticarte.com.au/pdf/WheatDArTmapsVersion1.2.xls). These map fragments were combined into a consensus map in two ways; firstly using a simple least squares (SLS) method in which the sum of squares of difference between the distances between marker pairs in the various map fragments was minimised by iteratively adjusting map positions in a consensus map (this was achieved using the 'Solver' function in MS Excel). Secondly consensus maps were produced using package MetaQTL (Veyrieras, *et al.* 2007) which implements a weighted least squares method to combine multiple maps.

Subsequently (Crossa *et al.*, 2007) (the CIMMYT map) produced an integrated map based on four spring wheat mapping populations and a sub-set of the alignments supplied by Triticate. This integrated map is available from the GrainGenes website (http://wheat.pw.usda.gov/GG2/index.shtml).

The three approaches gave similar ordering of markers on the majority of chromosomes (Figure 5-17 to Figure 5-21), notable discrepancy was evident in 1A, 2A, 3A and 2D. The SLS method tended to elongate maps, and MetaQTL shortened maps, compared to the values

[1] Marker was not polymorphic in the panel or it was impossible to score the genotype due to technical problems.

suggested by Somers *et al.* (2004) based on SSRs (Table 5-8). The Crossa maps were often short, but this was simply because the number of markers reported was less than that in the other maps.

Unsurprisingly (since they are derived from the same raw data) metaQTL and SLS agreed more closely with each other than they did with the Crossa map. There was some evidence (higher R^2 in Table 5-9) that the metaQTL map agreed more closely with the Crossa map than did SLS.

Table 5-8. Estimates of chromosome length from four different wheat consensus maps.

Chromosome	cM (Somers, 2004 – SSR)	cM (SLS - DArT)	cM (MetaQTL – DArT)	Crossa et al.(2007) (CIMMYT)
1A	126	166	140	102
1B	111	177	75	131
1D	117	129	115	133
2A	143	157	122	85
2B	123	222	129	117
2D	107	100	66	11
3A	116	185	112	49
3B	148	169	26	138
3D	79	57	71	6
4A	88	121	46	184
4B	59	117	92	149
4D	91	44	50	21
5A	184	163	173	169
5B	173	163	127	210
5D	120	115	79	24
6A	156	146	147	143
6B	82	173	132	181
6D	110	72	15	-
7A	131	171	131	226
7B	151	184	86	179
7D	154	138	70	-
Average chromosome length	122	141	95	119

Table 5-9. R^2 between the three consensus maps

	CIMMYT	metaQTL	SLS
CIMMYT	1	0.356484	0.297179
metaQTL	0.356484	1	0.495707
SLS	0.297179	0.495707	1

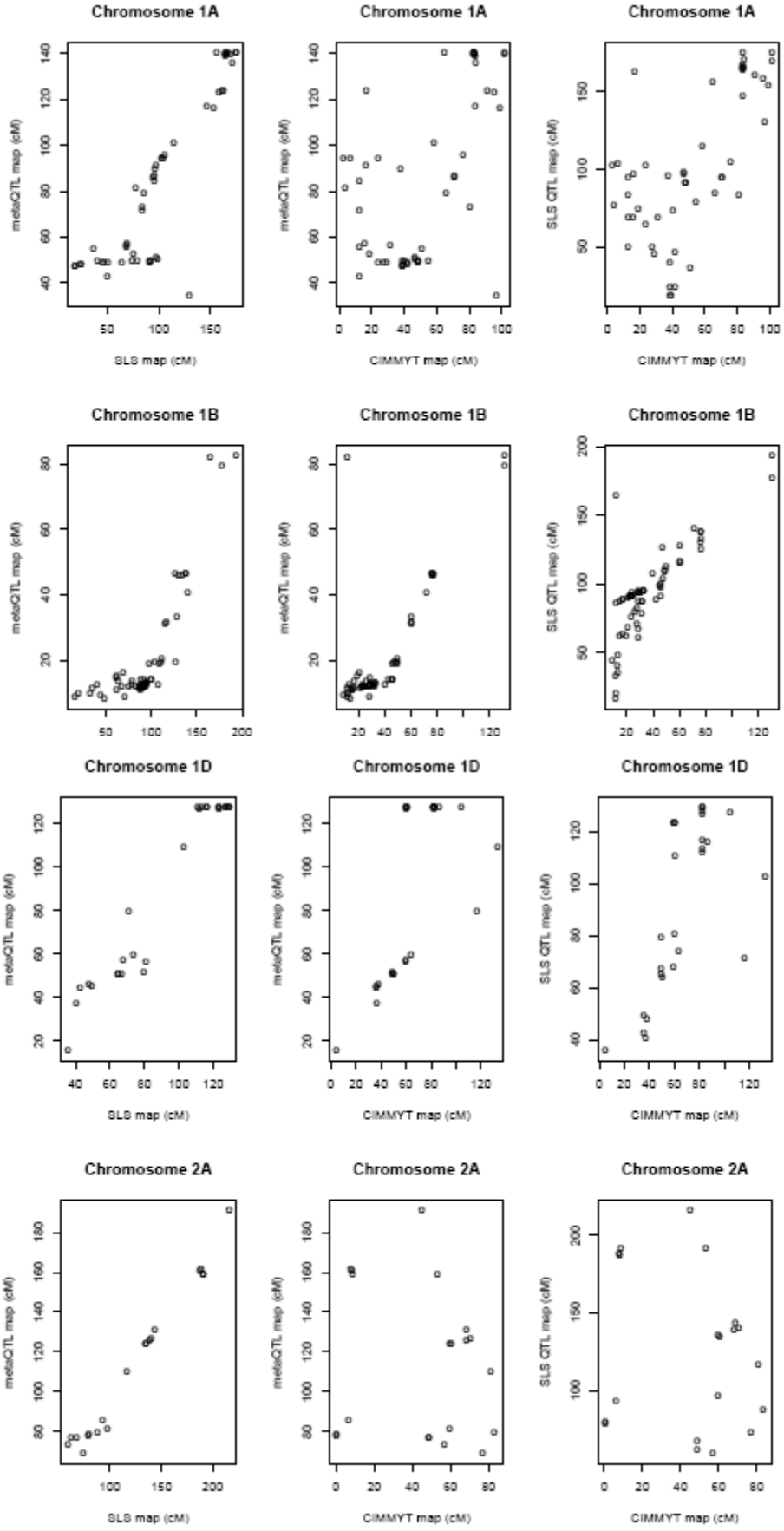


Figure 5-17. Chromosomes 1A, 1B, 1D and 2A. Comparisons of consensus maps obtained by Crossa et al. (CIMMYT map), metaQTL and Simple Least Squares (SLS).

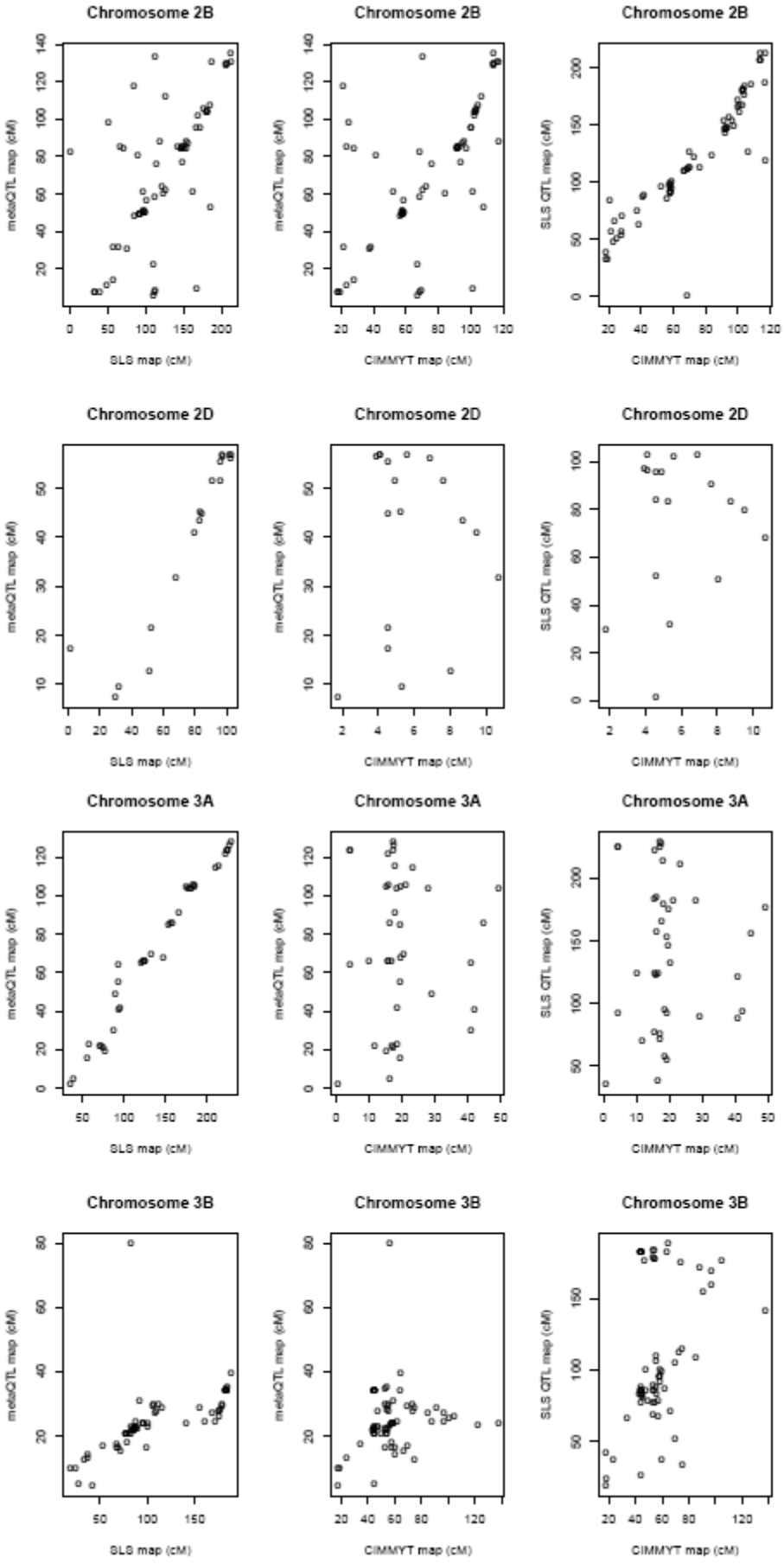


Figure 5-18. Chromosomes 2B, 2D , 3A and 3B. Comparisons of consensus maps obtained by Crossa et al. (CIMMYT map), metaQTL and Simple Least Squares (SLS).

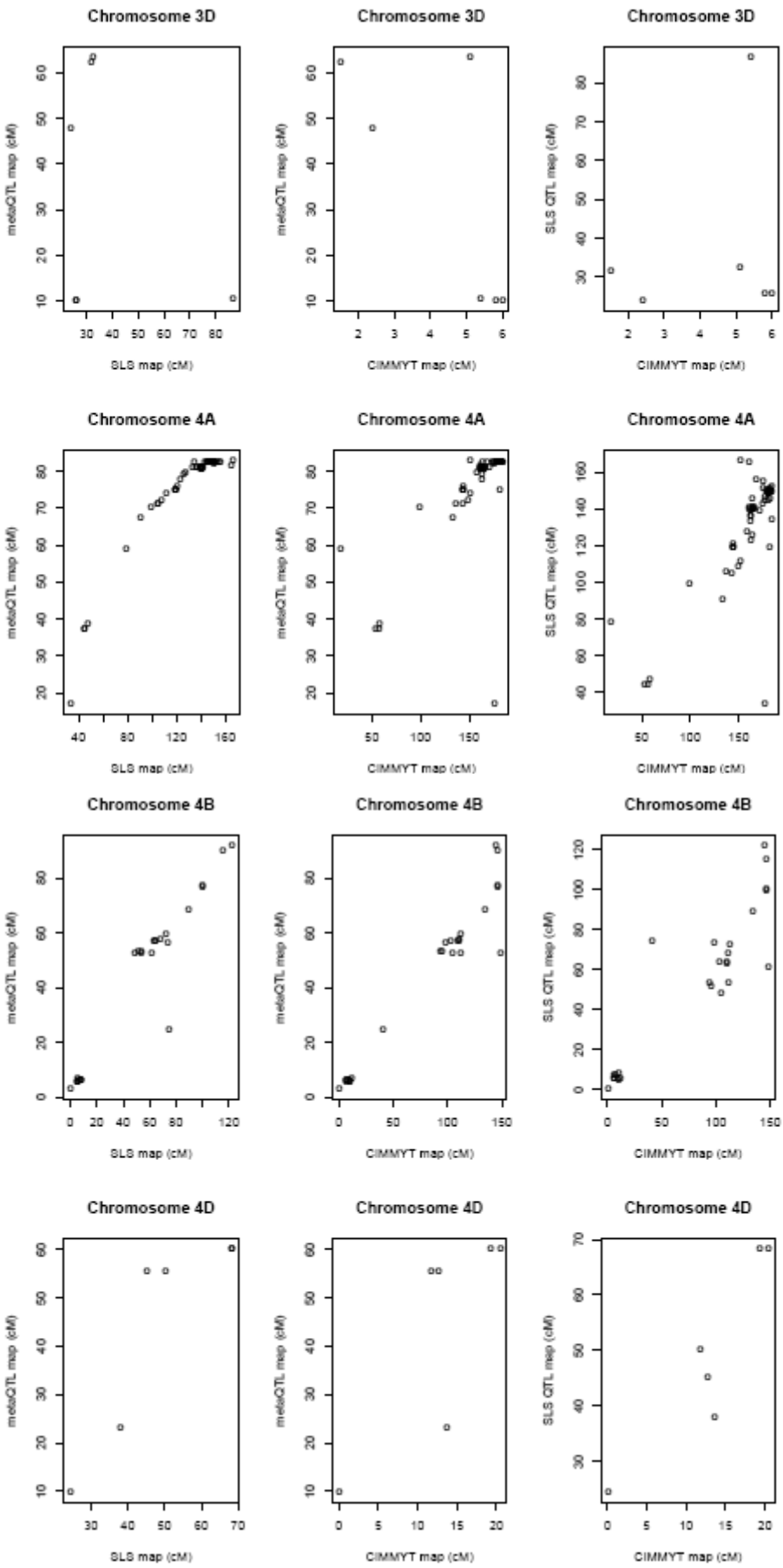


Figure 5-19. Chromosomes 3D, 4A, 4B and 4D. Comparisons of consensus maps obtained by Crossa et al. (CIMMYT map), metaQTL and Simple Least Squares (SLS).

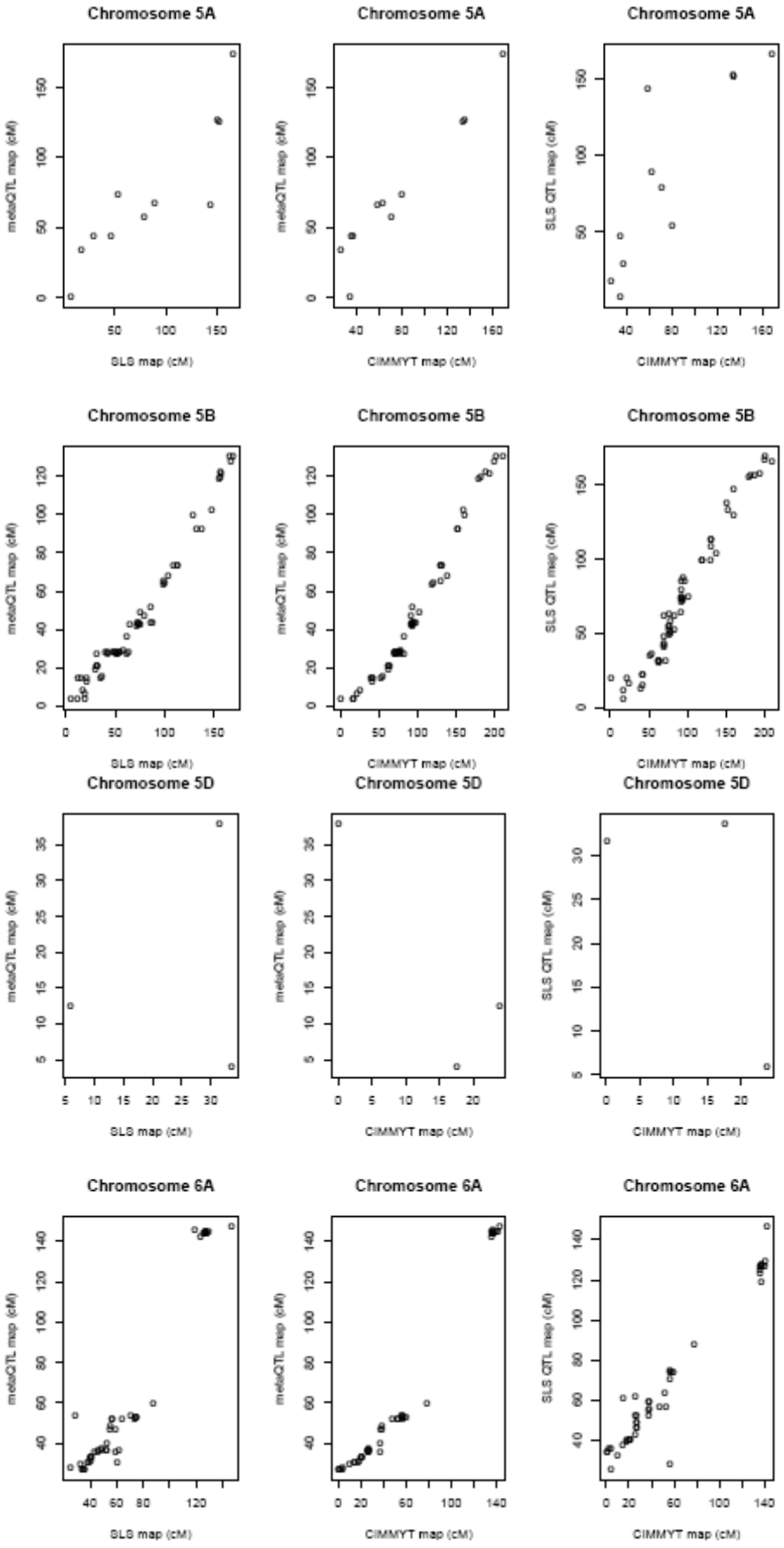


Figure 5-20. Chromosomes 5A, 5B, 5D and 6A. Comparisons of consensus maps obtained by Crossa et al. (CIMMYT map), metaQTL and Simple Least Squares (SLS).

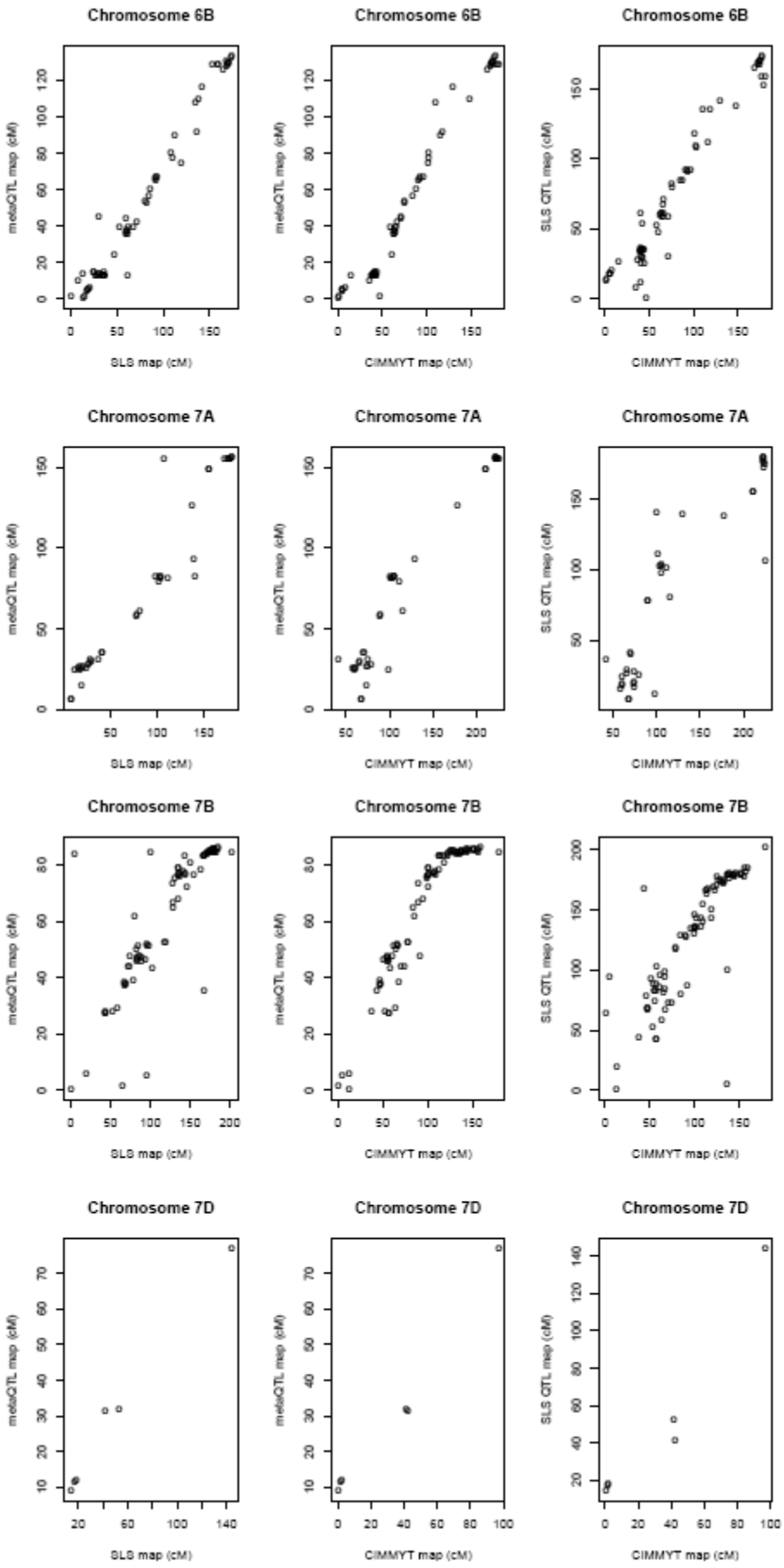


Figure 5-21. Chromosomes 6B, 7A, 7B and 7D. Comparisons of consensus maps obtained by Crossa et al. (CIMMYT map), metaQTL and Simple Least Squares (SLS). The absence of 6D is due to there being no CIMMYT integrated map for this chromosome.

5.2.4.1 DArT Map Choice.

It was evident that the SLS and MetaQTL (MET) maps were similar (as evidenced by the linear relationship between them in Figure 5-17 to Figure 5-21). Here onwards I simply considered the MET map and the CIMMYT map(CIM). In many cases the CIMMYT map was in good agreement with MET but in the case of chromosomes 1A, 2A, 3A and 2D the agreement was poor. CIM also failed to provide any mapped markers on chromosome 6D.

It was necessary to make an objective choice as to which map to use. Since the origins of the maps are estimates of recombination fraction between markers and since the use intended for the map was to attempt to estimate the genetic location of QTL the criterion for choosing between the maps should be an estimate of how well they model patterns of recombination in the data. In the absence of direct measurements of recombination fraction I calculated D' between marker pairs on each chromosome and used the *optim()* function in R Version 2.9.1 (2009) to develop a model for D' over genetic distance as defined by MET and CIM. The model was a least squares solution to the line of best fit through the data for the Malecot curve (Equation 1; Malecot 1948), which gives the theoretical relationship between D' and; the recombination fraction (θ) the asymtopic value of D'(l), D' at θ =0 (m) and the age of the population in meioses (t).

$$\hat{D}' = l + (m - l)e^{-\theta t}$$ Equation 1

Both MET and CIM are in Kosambi (1944) mapping units and were converted to recombination fractions using the inverse of the mapping function (Equation 2) which relates genetic distance (M) in Morgans and recombination fraction (θ).

$$\theta = \frac{1}{2}\frac{e^{4M} - 1}{e^{4M} + 1}$$ Equation 2

It was then possible to compare $\hat{D}'$ with observed D' using linear correlation and the properties of the residuals.

The significance of differences in correlation between the two models was tested by bootstrapping the observed data for each map 1000 times, the Malecot curve was fitted to each re-sampling and $\hat{D}'$ recalculated. From this the correlation coefficient for each pair of resampled data could be calculated and the distribution of the differences between them obtained. The distribution was centralised on zero by subtracting the mean to allow a test for H_0 ~ the difference in correlation coefficient is zero. Analyses were conducted on all mapped

markers, mapped markers common to both maps and data partitioned between chromosomes with good and poor between map agreement.

All available mapped markers from both CIM and MET

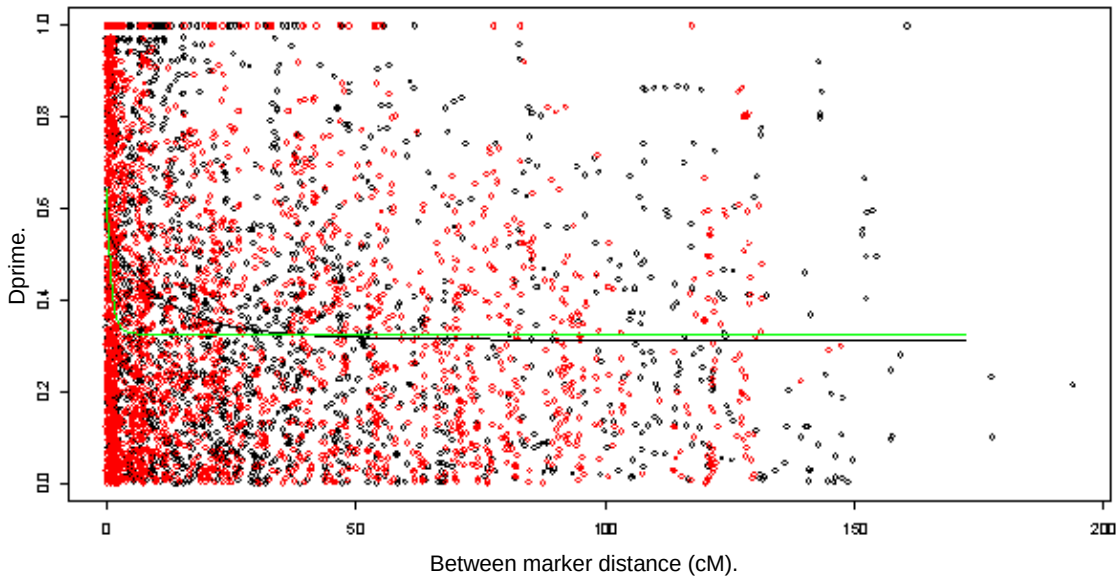


Figure 5-22. D' over between marker pair distance for all mapped markers. Lines are the best fit Malecot curves estimated by least squares. CIM (black circles, black line) and MET (red circles, green line).

Statistical analysis:

Variance of residuals ($\hat{D}'$ - D') (CIM):	0.085
Variance of residuals ($\hat{D}'$ - D') (MET):	0.082
P-value from F-test that variances are equal:	0.200.
Correlation of $\hat{D}'$ with observed D' (CIM):	0.254
Correlation of $\hat{D}'$ with observed D' (MET):	0.293
Difference in correlation:	0.039

P-value from t-test that correlation = 0: $<10^{-16}$ (MET and CIM)

Bootstrap result.

95% confidence interval for difference in correlation =0: -0.060-0.063
Empirically estimated p-value for H_0 ~ 'difference in correlation is 0' = 0.11

All mapped markers from Chromosomes 1A, 2A, 3A and 2D only.

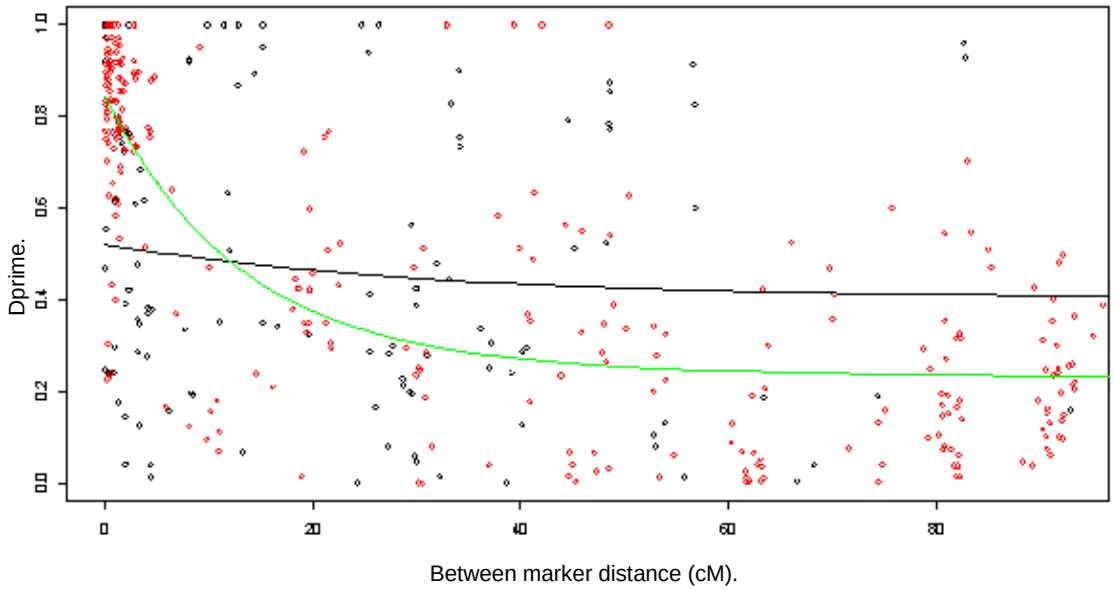


Figure 5-23. D' over between marker pair distance for all mapped markers on Chromosomes 1A, 2A, 3A and 2D. Lines are the best fit Malecot curves estimated by least squares. CIM (black circles, black line) and MET (red circles, green line).

Statistical analysis:

Variance of residuals ($\hat{D}'$ - D') (CIM):	0.101
Variance of residuals ($\hat{D}'$ - D') (MET):	0.041
P-value from F-test that variances are equal:	$9x10^{-10}$
Correlation of $\hat{D}'$ with observed D' (CIM):	0.113
Correlation of $\hat{D}'$ with observed D' (MET):	0.783
Difference in correlation:	0.670

P-value from t-test that correlation = 0: $<10^{-16}$(MET), 0.235 (CIM)

Bootstrap result.
95% confidence interval for difference in correlation =0: -0.157-0.194
Empirically estimated p-value for H_0 ~ 'difference in correlation is 0' < 0.001

All mapped markers common to CIM and MET

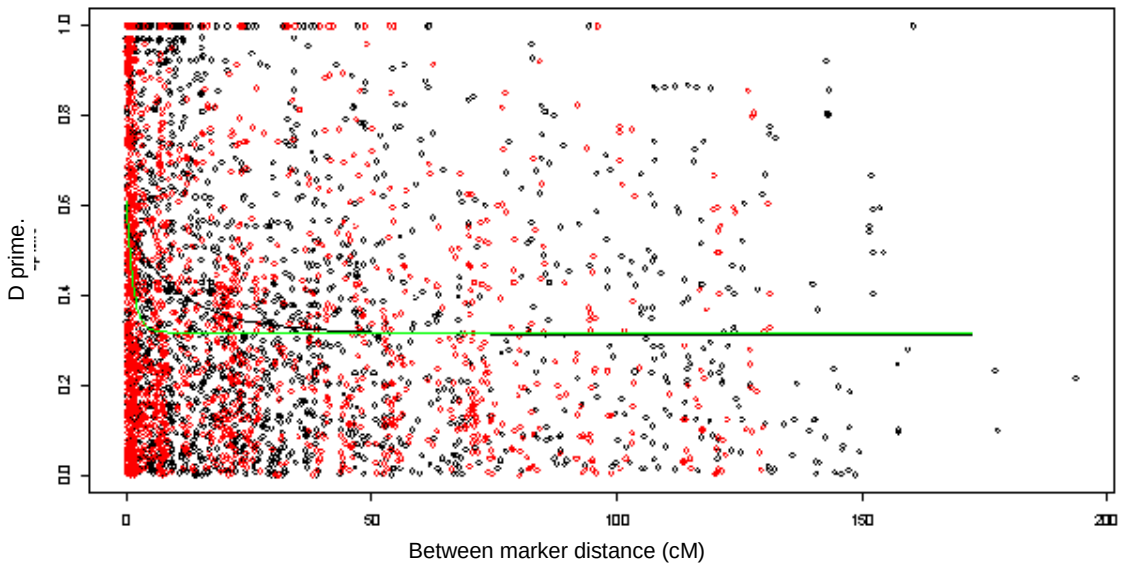


Figure 5-24. D' over between marker pair distance mapped markers common to CIM and MET. Lines are the best fit Malecot curves estimated by least squares. CIM (black circles, black line) and MET (red circles, green line).

Statistical analysis:

Variance of residuals ($\hat{D}'$ - D') (CIM):	0.085
Variance of residuals ($\hat{D}'$ - D') (MET):	0.081
P-value from F-test that variances are equal:	0.143
Correlation of $\hat{D}'$ with observed D' (CIM):	0.255
Correlation of $\hat{D}'$ with observed D' (MET):	0.308
Difference in correlation:	0.053

P-value from t-test that correlation = 0: $<10^{-16}$ (MET and CIM)

Bootstrap result.

95% confidence interval for difference in correlation =0: -0.062-0.066
Empirically estimated p-value for H_0 ~ 'difference in correlation is 0' = 0.055

Markers common to both maps from Chromosomes 1A, 2A, 3A and 2D only.

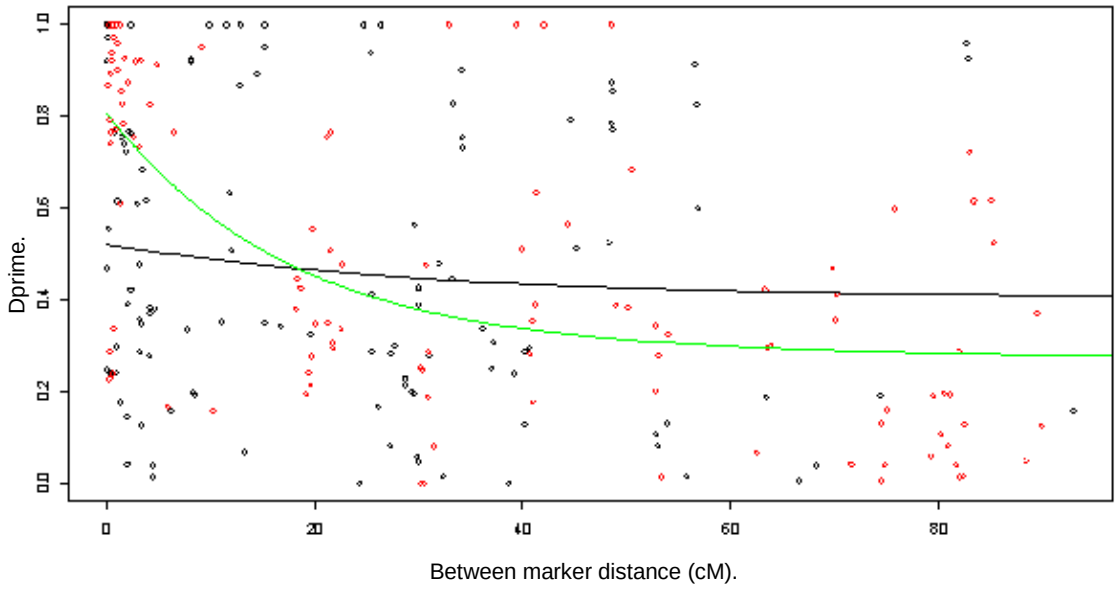


Figure 5-25. D' over between marker pair distance for all mapped markers on Chromosomes 1A, 2A, 3A and 2D common to both maps. Lines are the best fit Malecot curves estimated by least squares. CIM (black circles, black line) and MET (red circles, green line).

Statistical analysis:

Variance of residuals ($\hat{D}'$ - D') (CIM):	0.101
Variance of residuals ($\hat{D}'$ - D') (MET):	0.061
P-value from F-test that variances are equal:	0.004
Correlation of $\hat{D}'$ with observed D' (CIM):	0.113
Correlation of $\hat{D}'$ with observed D' (MET):	0.632
Difference in correlation:	0.519

P-value from t-test that correlation = 0: $<10^{-14}$ (MET). 0.235 (CIM).

Bootstrap result.

95% confidence interval for absolute difference in correlation =0: -0.217-0.233
Empirically estimated p-value for H_0 ~ 'difference in correlation is 0' < 0.001

Markers common to both maps from chromosomes other than 1A,2A,3A and 2D.

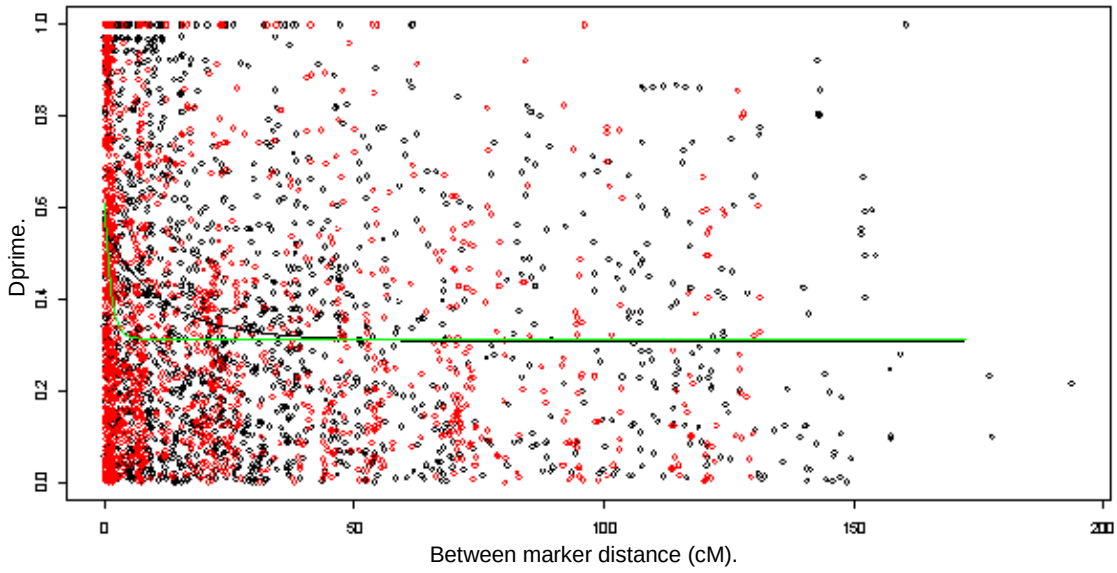


Figure 5-26. *D' over between marker pair distance mapped markers common to CIM and MET. Lines are the best fit Malecot curves estimated by least squares. CIM (black circles, black line) and MET (red circles, green line).*

Statistical analysis:

Variance of residuals ($\hat{D}'$ - *D'*) (CIM):	0.084
Variance of residuals ($\hat{D}'$ - *D'*) (MET):	0.080
P-value from F-test that variances are equal:	0.195
Correlation of $\hat{D}'$ with observed *D'* (CIM):	0.261
Correlation of $\hat{D}'$ with observed *D'* (MET):	0.313
Difference in correlation:	0.052

P-value from t-test that correlation = 0: $<10^{-16}$ (CIM and MET)

Bootstrap result.
95% confidence interval for absolute difference in correlation =0: -0.068-0.070
Empirically estimated p-value for H_0 ~ 'difference in correlation is 0' =0.07

Final map choice.

Where comparisons between CIM and MET were made in terms of the correlation of $\hat{D}'$ and D', irrespective of the marker sub-sets used the model based on MET was always more

highly correlated. This difference in the correlation was significant ($P<0.05$) when the models were developed on data only from chromosomes 1A,2A,3A and 3D, this was true for all available markers and also when the data were partitioned to include only all common markers. The difference was also significant ($P<0.1$) when the model was developed on all common markers and for all common markers from chromosomes other than 1A,2A,3A and 2D.

The difference in correlation between models when all available markers over all chromosomes were used was not significant ($P>0.1$) although the empirical p-value of 0.11 makes this decision somewhat borderline.

For the markers on chromosomes 1A,2A,3A and 2D, whether using all markers or just common markers, the variances of residuals for CIM and MET were not equal ($P<0.01$) and CIM had the higher variance.

These analyses suggest that MET was a better map, particularly for the four chromosomes identified as anomalous because it allowed a better model of decay of D' to be constructed. When the genome was considered as a whole, models of the decay of D' based on MET showed significantly higher correlation with observed D'. This might be simply the effect of the four anomalous chromosomes but the difference in correlation between the models persisted, with modest significance, even if these chromosomes were excluded.

In addition to describing a better model of LD decay MET had the advantage of containing more DArT markers together with a range of SSR, AFLP and BARC markers which could be useful for cross-referencing to other genetic maps.

I concluded that MET should be the map used in this study.

5.2.5 Population Sub-structure.

The wheat panel was selected from amongst the pool of winter sown wheat varieties, this avoided one potential source of population sub-structure (see; Barley for comparison 5.1.4). However, the winter wheat crop has been subject to plant breeding activity and includes, for example, divisions between wheat destined for bread-making, confectionary flour and animal feeding.

Examining the first two principal coordinates of the genetic correlation between individuals in the wheat panel (Figure 5-27) showed a clear division in the population, this was attributable to the introgression of a fragment of chromosome 1R from rye into wheat chromosome 1B. This introgression brought useful disease resistance but is generally associated with poor milling quality.

The presence of highly correlated groups of markers can distort models of sub-structure, it was found that, unlike barley (Figure 5-4) the removal of one from each highly correlated marker pair reduced considerably the clarity of the division (Figure 5-28). This was reflected in a fall in F_{ST} from 0.303 to 0.068 between the two datasets.

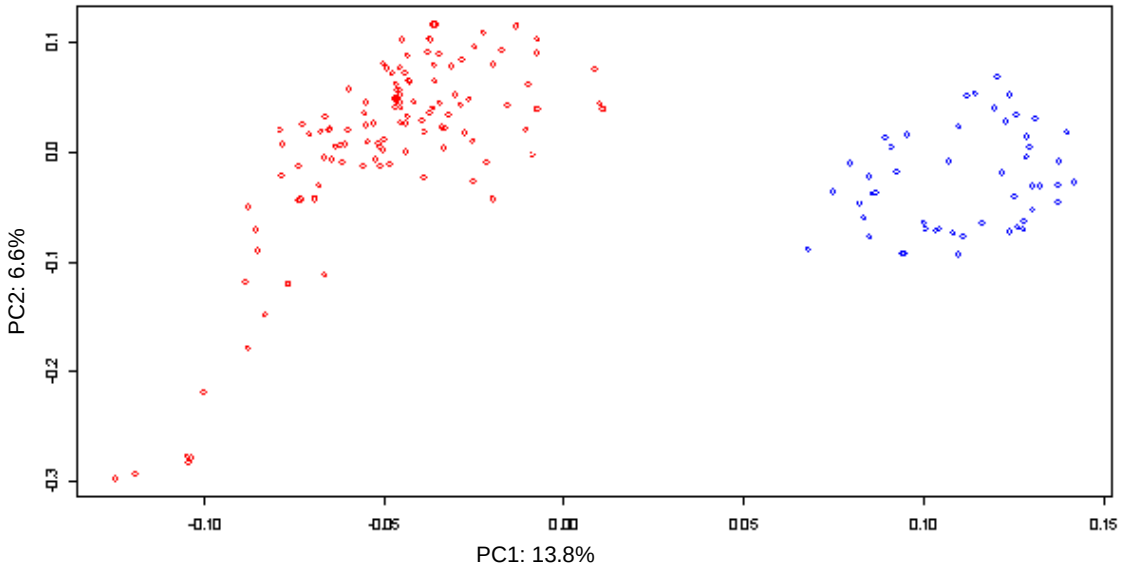


Figure 5-27. First two principal coordinates of a genetic correlation matrix between individuals using all available markers (MAF>0.1). Red circles represent wheat varieties which do not contain the 1B1R translocation, blue circles contain the translocation.

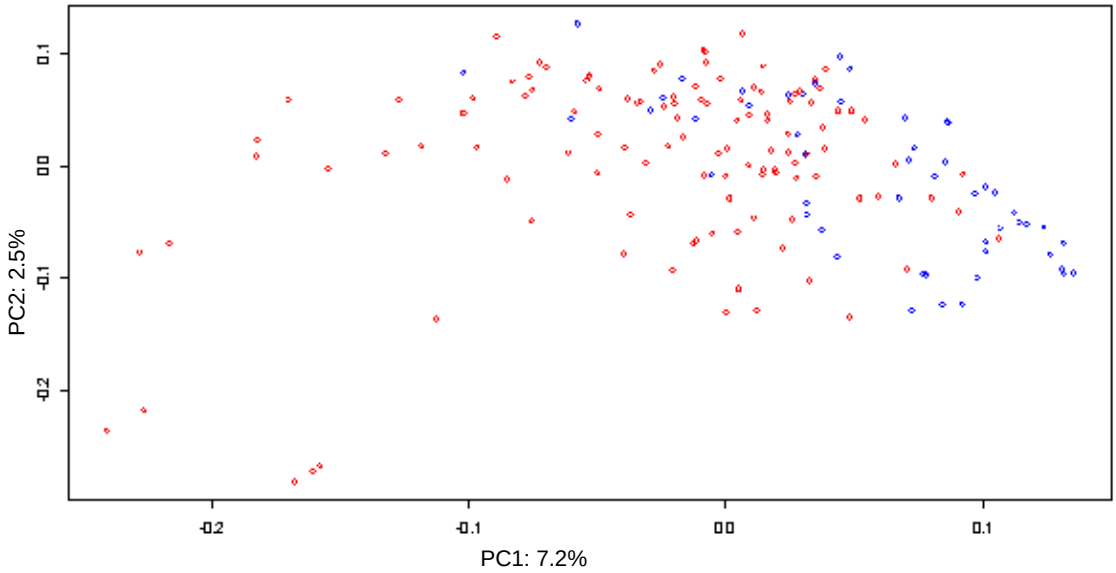


Figure 5-28. First two principal coordinates of a genetic correlation matrix between individuals using only uncorrelated (r^2<0.6) markers from among all available markers (MAF>0.1). Red circles represent wheat varieties which do not contain the 1B1R translocation, blue circles contain the translocation.

It was possible to be more specific in identifying the source of the division between these sub-populations if mapped markers were used. Where all mapped markers were used to estimate genetic correlation the representation of structure was similar to that seen with all available markers (Figure 5-27) and F_{ST} = 0.25. Removing one of each highly correlated marker pair had the effect of reducing F_{ST} to 0.054 and, as previously, the principal component analysis showed little evidence of structure (Figure 5-30).

Finally, the removal of all markers mapped to chromosome 1B, but including all other mapped markers (even highly correlated pairs), showed that the subpopulation structure was being defined by a few markers on chromosome 1B since their removal reduced F_{ST} to 0.065 and no substructure was evident in the first two principal coordinates (Figure 5-31).

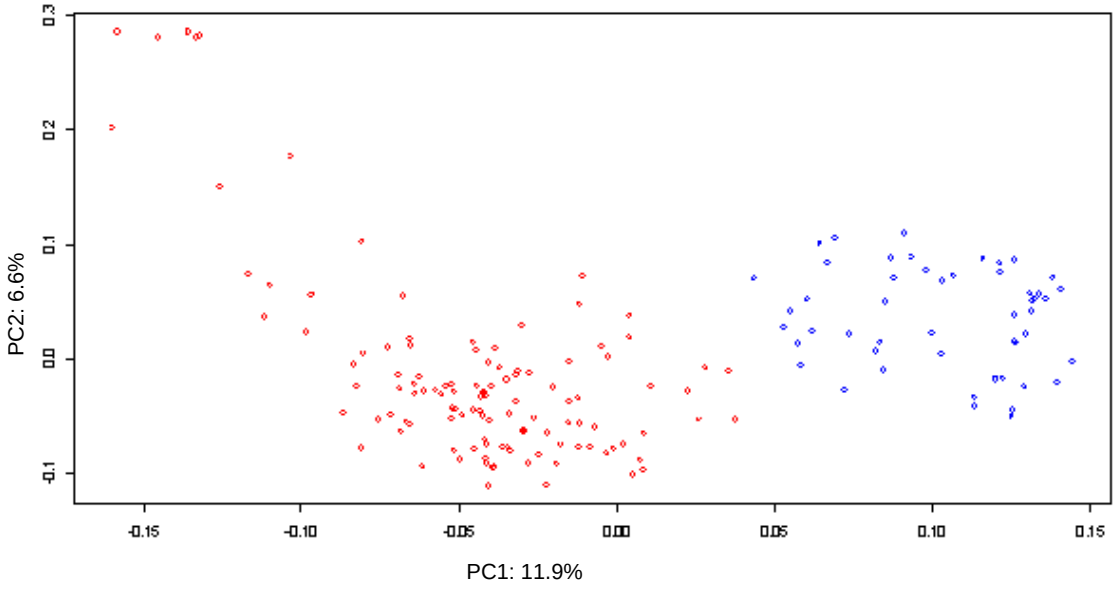


Figure 5-29. First two principal coordinates of a genetic correlation matrix between individuals using all mapped markers (MAF>0.1). Red circles represent wheat varieties which do not contain the 1B1R translocation, blue circles contain the translocation.

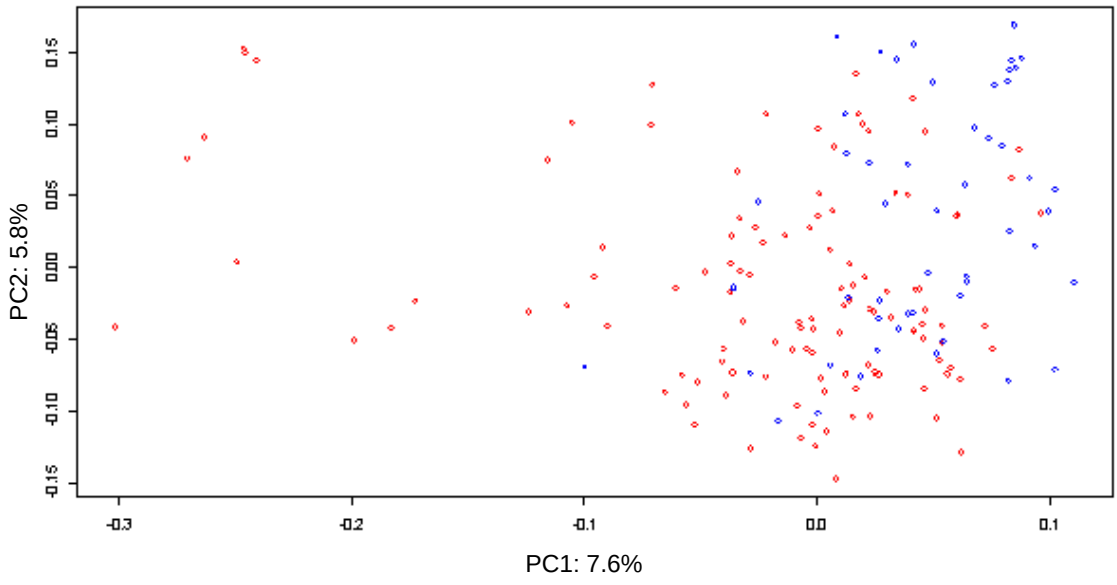


Figure 5-30. First two principal coordinates of a genetic correlation matrix between individuals using only uncorrelated (r^2<0.6) markers from among all mapped markers (MAF>0.1). Red circles represent wheat varieties which do not contain the 1B1R translocation, blue circles contain the translocation.

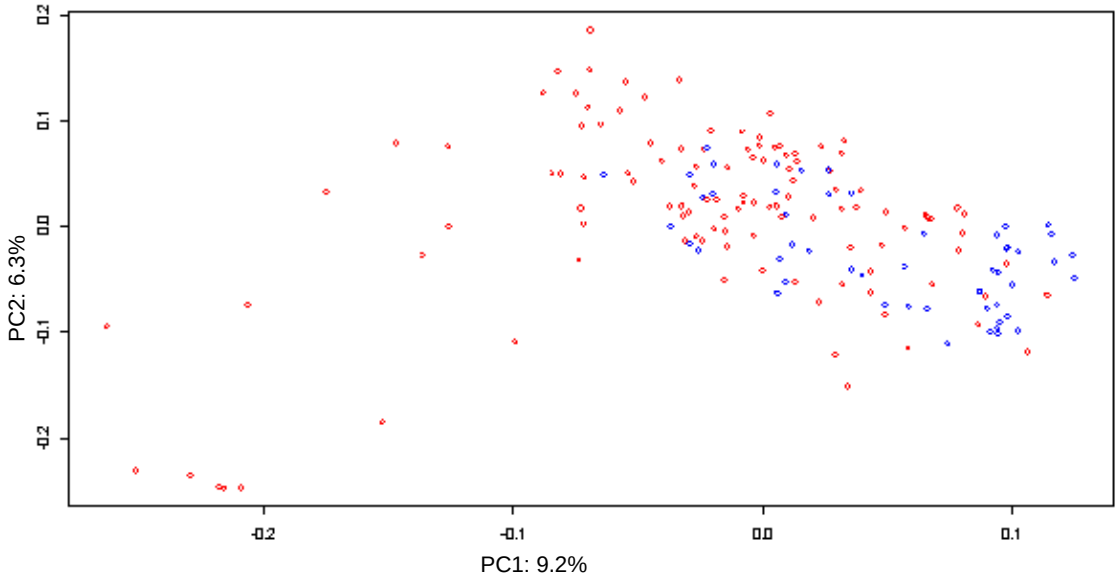


Figure 5-31. First two principal coordinates of a genetic correlation matrix between individuals using all mapped markers (MAF>0.1) except those mapping to chromosome 1B. Red circles represent wheat varieties which do not contain the 1B1R translocation, blue circles contain the translocation.

5.3 Simulated Cereal Population.

The available wheat and barley panels were constrained in terms of numbers of individuals and, to an extent, marker panel size. These datasets would allow investigation of the statistical power of panels smaller than themselves but the potential gain in power from larger panels required simulated genetic data.

5.3.1 Simulated Population Development.

The program ‘MS’ (Hudson, R. R., 2002) was used to develop a simulated cereal population under the Wright-Fisher neutral model.

Memory constraints on the UNIX server limited the size of population, and the number of segregating sites. In order to obtain a population approximating to a cereal species whilst working with these constraints I was forced to impose some unusual features on the population. The history of the population development is summarised in Table 5-10 and represented graphically in Figure 5-32.

Table 5-10. Summary of the genetic/evolutionary history of the simulated cereal population.

Feature	Value	MS command
Population age since speciation	~3x 10^6 meioses	Convergence of exponential growth to one individual.
Population age since first division into sub-populations	441 meioses	-ej 0.015 4 1 -ej 0.015 3 1
Population age since second division into sub-populations	44 meioses	-ej 0.0015 2 1
Migration between sub-populations.	Population 1 and 2, 10 individuals in either direction per generation. Population 3 and 4, 100 individuals in either direction per generation. Population 3 to population 1, 10 individuals per generation. Population 4 to population 2, 10 individuals per generation.	-m 0 1 2 10 -m 0 2 1 10 -m 0 3 4 100 -m 0 4 3 100 -m 0 1 3 10 -m 0 2 4 10
Growth rate per *t* generations, constant throughout history of the population.	$e^{-0.06259t}$	-eG 0.000025 0.06259
Sudden population growth one meiosis before present.		-eN 0.000025 0.0225
Diploid population size at the end of the simulation (N_0)	Consists of four sub-populations each of size 3677.	-r 29416 24000[1]
Number of segregating sites	15698	-s 15698 [2]

[1] -r 29416 24000 sets the length of the simulated chromosome to be 24000 base pairs. The probability of a crossing over event is considered to be ~1 because the length of the simulated chromosome is 1720cM. Hence $4N_0 = 29416$ and $N_0 = 7354$.

[2] I found that allowing a population to develop over a long period with fixed mutation rate produced too few markers with MAF>0.1. To overcome this I forced a fixed number of segregating sites (15698, the maximum the UNIX server could accommodate) over an extended period of time (3x10^6 meioses). By fixing the number of segregating sites I ensured that no polymorphic site was lost due to drift; the software ensured a single mutation took place in each generation for each mono-morphic site. Nonetheless, only a small proportion of sites achieved a minimum allele frequency greater than 0.1.

Action at end of simulation	All subpopulations sampled to create panel of 2000 haploid individuals. Sampling population 1 & 2 (666 haploid individuals) 3 & 4 (334 haploid individuals). The resulting simulated chromosome has a notional physical length of 1 arbitrary unit with the 15698 segregating sites distributed over its length. The scale was converted to notional cM by multiplying by 1720cM (this assumes a linear relationship between physical and genetic distance). The chromosome was then broken into seven 160cM chromosomes by removing six 100cM segments. Comparison of markers to marker LD and association between the segements showed that association was independent of the original distance between the markers and the segments were regarded as separate chromosomes. Segregating sites with MAF<0.1 were discarded.	-I 334 334 666 666

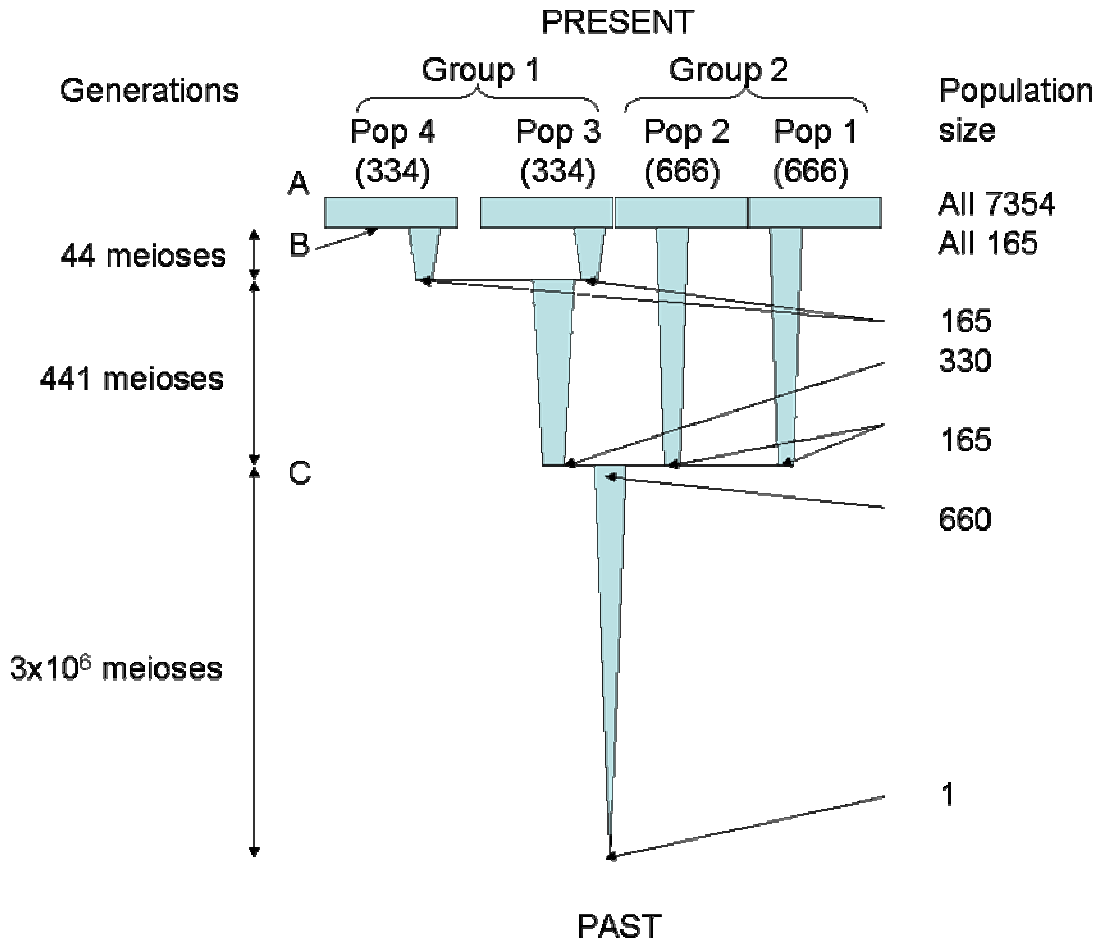


Figure 5-32. Diagrammatic representation of the history of the simulated cereal population. Numbers in brackets are the sample size taken from the four sub-populations to form the final panel.

The structural and LD characteristics of the population were imposed, firstly by random mating over a long ($3x10^6$ generation) burnin period. Since a fixed number of segregating sites was maintained through this period the number of sites achieving useful allele frequencies (MAF>0.1) was higher than would have arisen using a natural mutation rate. This approach was necessary to overcome the limitations imposed by the computing power available which in turn forced the simulated genome segment to be short in comparison to size of a natural genome. Had it been possible to simulate an entire genome then mutation and drift would have provided a large number of useful polymorphisms, although these would have been a very small fraction of all the potential mutation sites. Following this burnin Figure 5-33 shows that at point C in Figure 5-32 a population has been established which shows a clear LD decay curve over genetic distance. Random mating has prevented the formation of sub-population structure, this is evident visually in (Figure 5-34) and can

also be demonstrated statistically; for example AMOVA (Excoffier, 1992) can be used to demonstrate that the genetic variation in the population (represented by an Euclidian distance matrix between individuals) is coming solely from the between haplotype (individuals) variation (Table 5-11) and that F_{ST} (the fixation index; Wright (1951,1965)) is approximately zero (see;**Table 5-11** for detail).

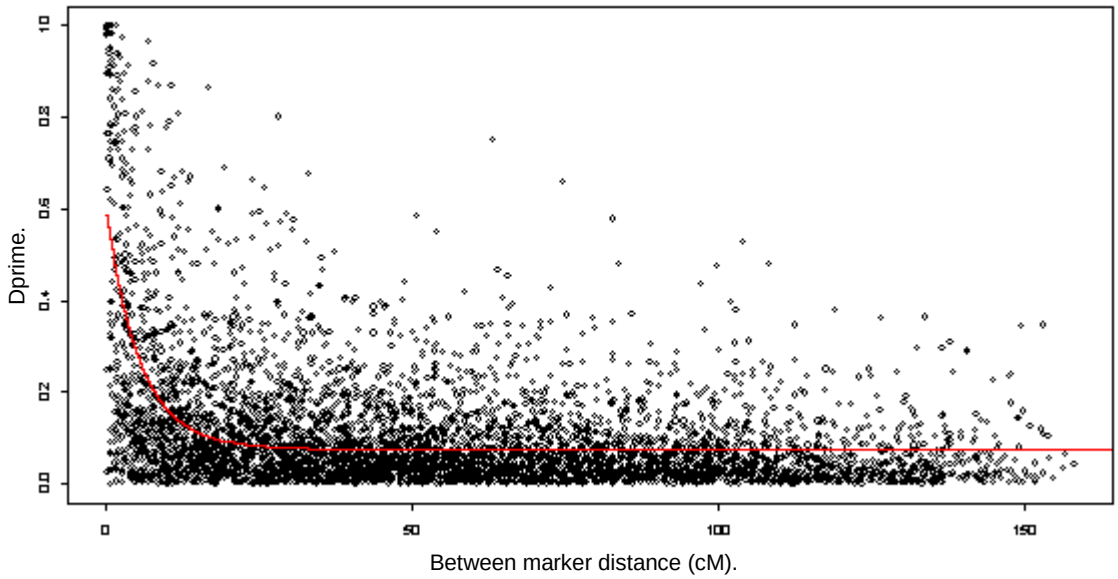


Figure 5-33. LD decay profile at point C in Figure 5-32. LD decays over distance and high values of D' become rare beyond ~ 80cM. However, LD has not fully decayed due to the continuous introduction of new mutations at a rate far in excess of that which would occur in nature.

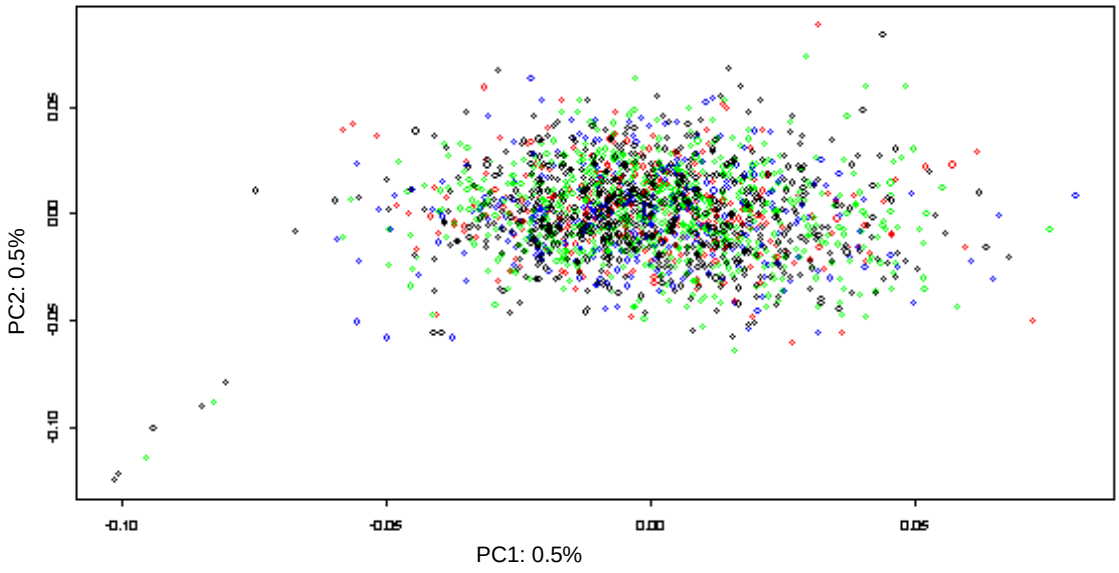


Figure 5-34. First two principal coordinates of the genetic correlation matrix of a sub-set of moderately uncorrelated (r^2<0.6) markers. Sample taken at point C in Figure 5-32. At this stage in population development there are no sup-populations. However the panel has been divided into four notional sub-populations according to the sampling strategy used throughout. This serves to illustrate the absence of obvious structure. Pop1 = black, pop 2 = green, pop 3 = red, pop4 = blue.

Table 5-11. AMOVA Merged populations at point C

Model: Euclidian distance matrix ~ groups/populations.

Source of variation	SSD	MSD	df	Expected MS
Among groups	461.2531	461.2531	1	$n''\sigma_a^2 + n'\sigma_b^2 + \sigma_c^2$
Among pops/within groups	983.9079	491.954	2	$n\sigma_b^2 + \sigma_c^2$
Within pops	959468.2	480.9364	1995	σ_c^2
Total	960913.3	480.9376	1998	σ_T^2

Components of covariance	σ^{21}	%[1]	p-value
σ_a^2	-0.033	-0.0069	1.00[2]
σ_b^2	0.022	0.004	0.28[3]
σ_c^2	480.94	100.002	0.46[4]
σ_T^2	480.93	100.0	

$\sigma_a^2, \sigma_b^2, \sigma_c^2 =$ variance components of (respectively) the group, the population and individual haplotype effects in the underlying model;

$$x_{ijk} = x + a_k + b_{jk} + c_{ijk}$$

$F_{ST} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_T^2} = 0.00$ where x is the ith haplotype frequency vector for the jth population in the ith group.

Between points C and B (Figure 5-32) sub-populations are allowed to develop with limited migration. The process of drift leads to differing allele frequencies at many loci so that although each population exhibits LD decay within the length of a chromosome merging the populations together at point B (Figure 5-32) would yield a population with sub-populations

[1] Small negative values for σ^2 and consequently % result from fitting an AMOVA model to data which contains no structure, true values are expected to be 0.
[2] H_1 ~ Observed covariance is greater than the covariance under the null (distribution under null estimated from 1000 permutations).
[3] H_1 ~ Observed covariance is greater than the covariance under the null (distribution under null estimated from 1000 permutations).
[4] H_1 ~ Observed covariance is less than the covariance under the null (distribution under null estimated from 1000 permutations).

visible in the PCA plot (Figure 5-35 and Figure 5-36). AMOVA analysis at point B (**Table 5-12**) shows about 1.6% of the variation is explained by variation among groups and 0.4% by populations within groups. The estimate of F_{ST} = 0.02.

Table 5-12 AMOVA at point B in the development of the simulated cereal.

Source of variation	SSD	MSD	df
Among groups	4876.0	4876.0	1
Among pops/within groups	1766.0	883.0	2
Within pops	573926.3	287.7	1995
Total	580568.2	290.6	1998

Components of covariance	σ^2	%	p-value
σ_a^2	4.56	1.6	0.165[1]
σ_b^2	1.19	0.4	<0.001[2]
σ_c^2	287.68	98.0	<0.001[3]
σ_T^2	293.44	100.0	

$$F_{ST} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_T^2} = 0.0196$$

[1] H_1 ~ Observed covariance is greater than the covariance under the null (distribution under null estimated from 1000 permutations).
[2] H_1 ~ Observed covariance is greater than the covariance under the null (distribution under null estimated from 1000 permutations).
[3] H_1 ~ Observed covariance is less than the covariance under the null (distribution under null estimated from 1000 permutations).

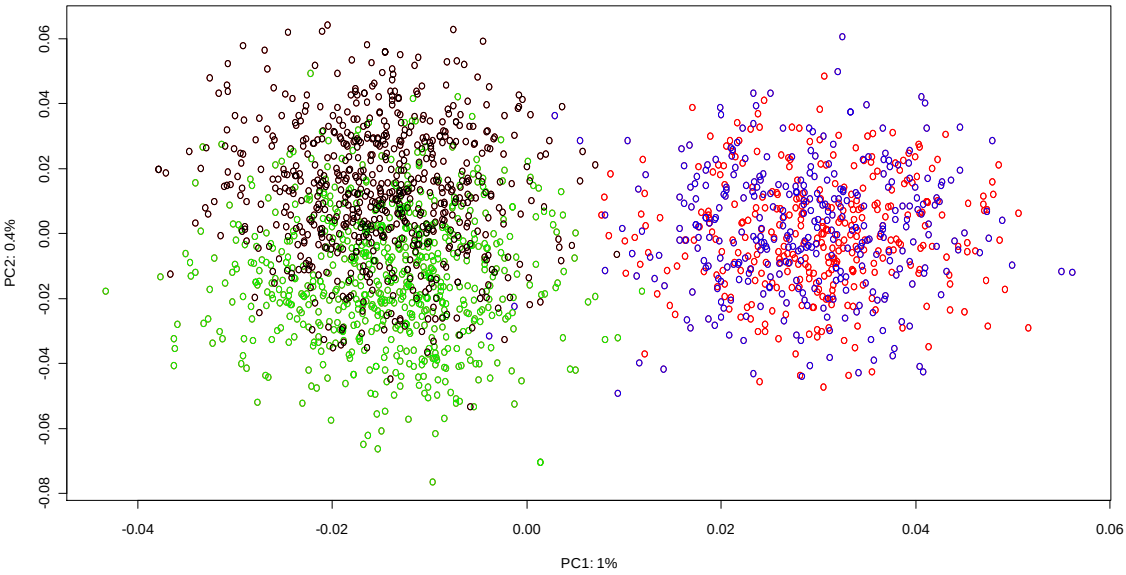


Figure 5-35. First two principal coordinates of the genetic correlation matrix of a sub-set of moderately uncorrelated (r^2<0.6) markers taken at point B B (Figure 5-32) in the development of the simulated cereal. Populations 1 and 2 and 3 and 4 are separating due to drift. The high migration rate between population 3 and 4 has prevented these from differentiating but lower migration rate between 1 and 2 has allowed some differentiation to occur. Pop1 = black, pop 2 = green, pop 3 = red, pop4 = blue.

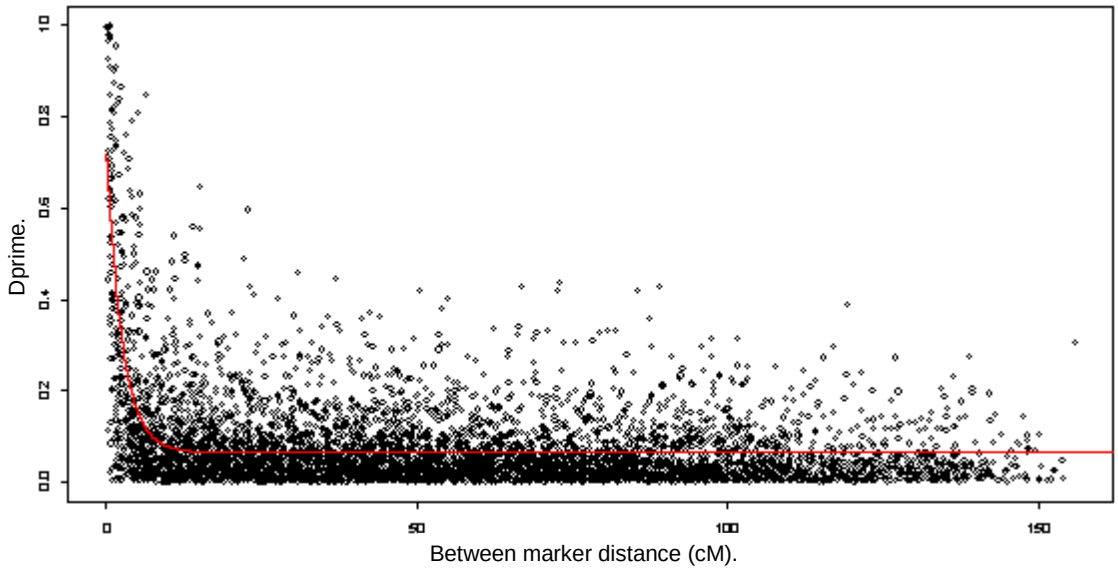


Figure 5-36. Decay of LD (as D') over between marker genetic distance in a sample taken at point B (Figure 5-32) in the development of the simulated cereal. All possible on chromosome pair-wise comparisons were made and then re-sampled with replacement to ensure the frequency of D' observations were evenly distributed over distance.

Between points B and A the sub-population sizes were all increased to 7354 haploid individuals. This was achieved, not by breeding over generations but by resampling within the sub-populations with replacement. Subsequently one further round of random mating took place within each subpopulation.

Finally at point A the panel was recovered by sampling each sub-popualtion as indicated in Figure 5-32 to give a total of 2000 individuals. This sub-sampling step imposed a bottleneck and sampling variation served to strengthen the between population and between group differences. Individul sub populations exhibit LD which show substantial decay within the length of a chromosome (Figure 5-37) however when these were merged the resultant final panel showed high LD extending over the full length of the chromosome and visually apparent structure in the PCA plot (see Figure 5-38 and Figure 5-39). AMOVA (Table 5-13) shows 45% of variation is explained by group and 9% by population within group. F_{ST} has risen to 0.54.

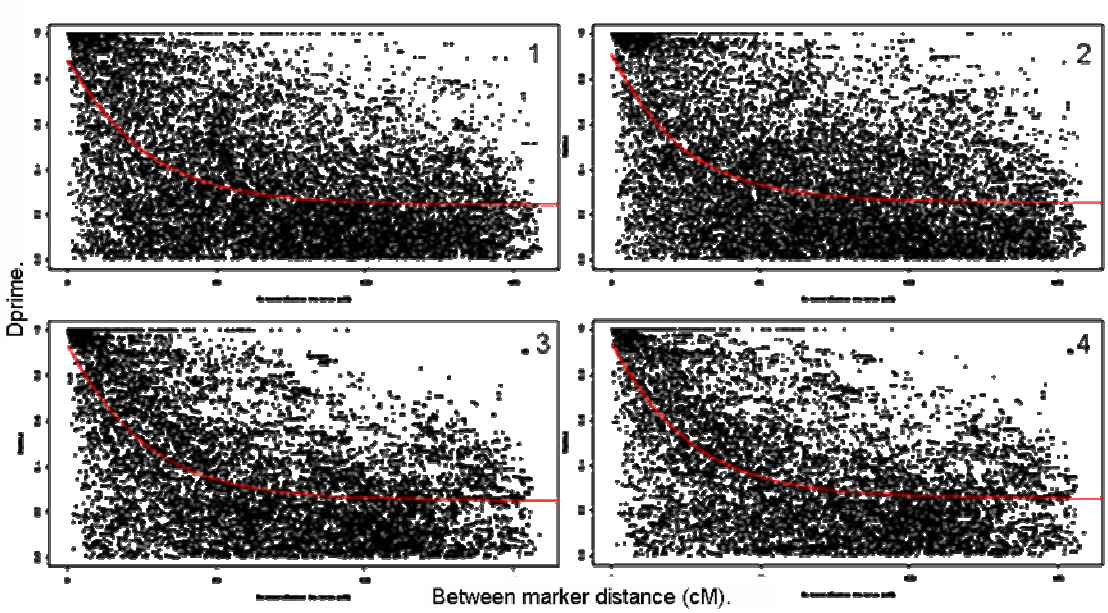


Figure 5-37. LD decay curves for the four sub-populations merged at point A in Figure 5-32. All possible on chromosome pair-wise comparisons were made and then re-sampled with replacement to ensure the frequency of D' observations were evenly distributed over distance. The red line is the Malecot curve through the data suggesting LD showing the asymptotic value of D' is lower in the sub-populations than in the composite population (Figure 5-39).

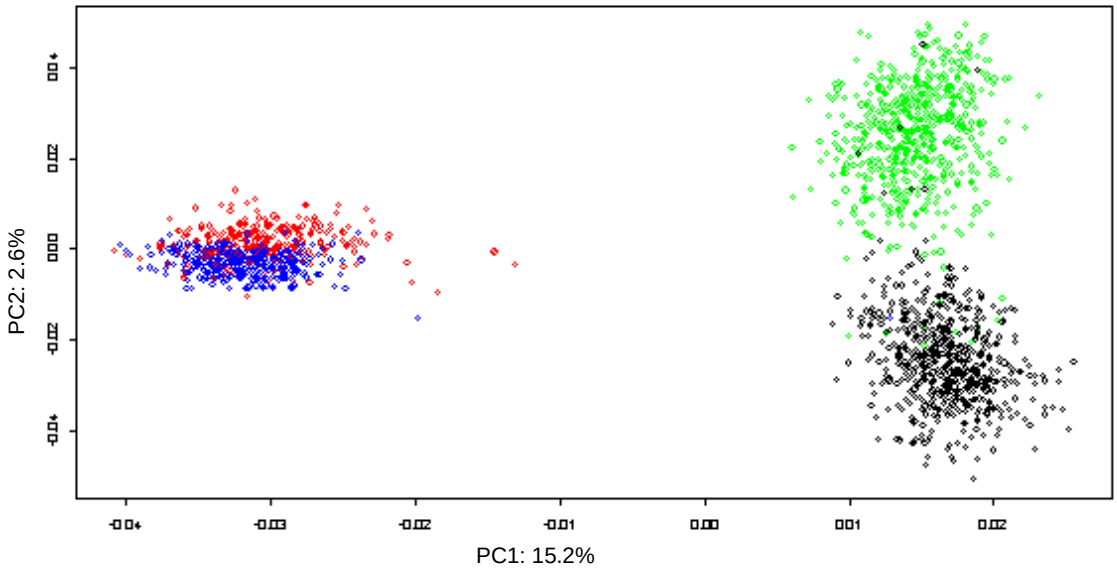


Figure 5-38. First two principal coordinates of the genetic correlation matrix of a sub-set of moderately uncorrelated (r^2<0.6) markers from the simulated cereal sampled at point A. The population bottleneck introduced by the final sampling event served to differentiate the sub-populations from each other and so emphasise the sub-population structure. Pop1 = black, pop 2 = green, pop 3 = red, pop4 = blue.

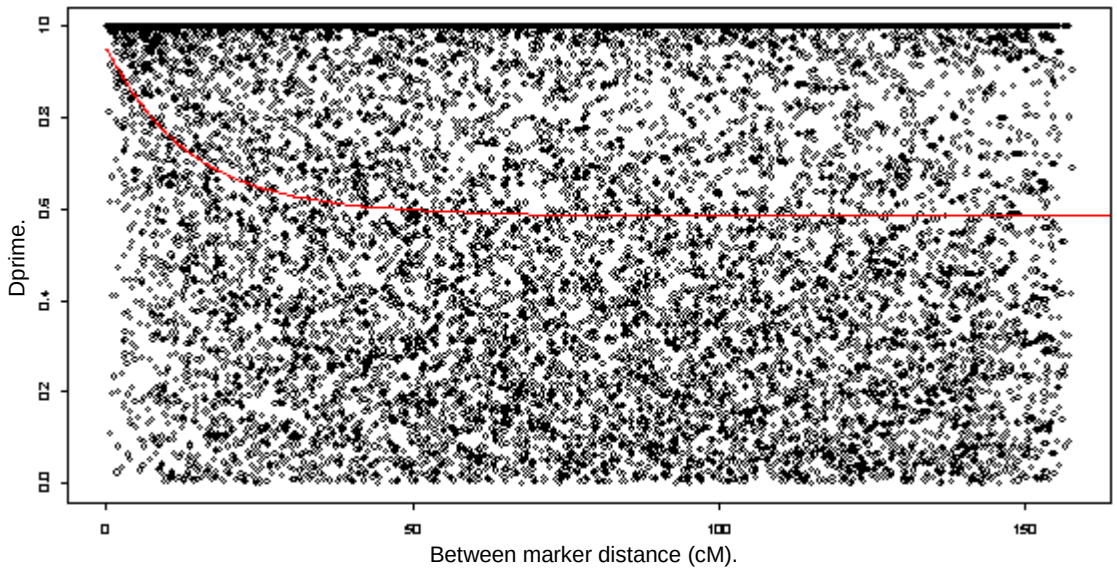


Figure 5-39. Simulated cereal. Decay of LD (as D') over between marker genetic distance. Composite of samples taken at point A in Figure 5-32. All possible on chromosome pair-wise comparisons were made and then re-sampled with replacement to ensure the frequency of D' observations were evenly distributed over distance. The red line is the Malecot curve through the data suggesting LD is elevated, but constant, even at large genetic distances. It is particularly noteable that high LD extends over the full length of the chromosome, this is explicable in terms the mingling of differentiated sub-populations in the final panel.

Table 5-13. AMOVA at point A in the development of the simulated cereal.

Source of variation	SSD	MSD	df
Among groups	148500.7	148500.7	1
Among pops/within groups	29285.7	14642.9	2
Within pops	310199.8	155.5	1995
Total	487986.2	244.2	1998

Components of covariance	σ^2	%	p-value
σ_a^2	152.3	45.2	<0.001[1]
σ_b^2	29.0	8.6	<0.001[2]
σ_c^2	155.5	46.2	<0.001[3]
σ_T^2	336.74702	100.0	

$$F_{ST} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_T^2} = 0.54$$

The genotypes produced underwent further processing after the simulation was complete. Specifically markers with MAF<0.1 were eliminated and the full simulated genome was broken into 7 chromosomes by the excision of 6 100cM segments. The intention being that the resulting 'chromosomes' would exhibit no trend in LD over the original distance between the segments, that is to say they would exhibit the characteristics of independent assortment within the limitations imposed by the population sub-structure.

Figure 5-40 and Figure 5-41 respectively show; the Malecot curve for the final population is asymptotic for between marker distances corresponding to the original distance between chromosomes in the simulated genome and that this effect can duplicated by fitting a Malecot curve through data representing the LD decay between markers on adjacent chromosomes

[1] H_1 ~ Observed covariance is greater than the covariance under the null (distribution under null estimated from 1000 permutations).
[2] H_1 ~ Observed covariance is greater than the covariance under the null (distribution under null estimated from 1000 permutations).
[3] H_1 ~ Observed covariance is less than the covariance under the null (distribution under null estimated from 1000 permutations).

over the original distance between them.

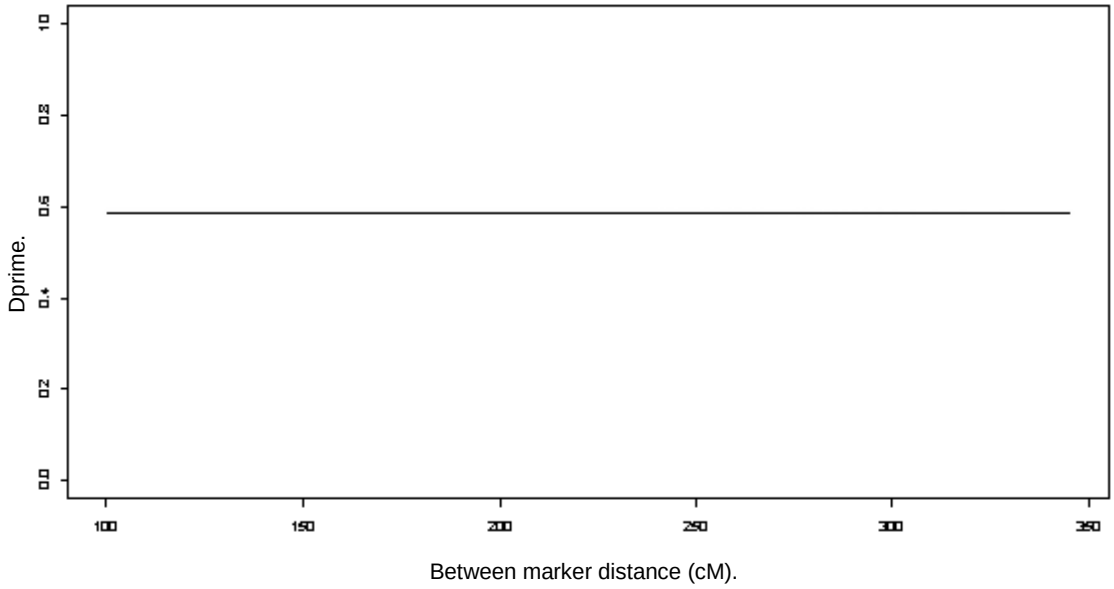


Figure 5-40. The Malecot curve for the merged population at point A is asymptotic at >100cM. Asymptote = 0.587. This suggests adjacent chromosomes will be assort independently.

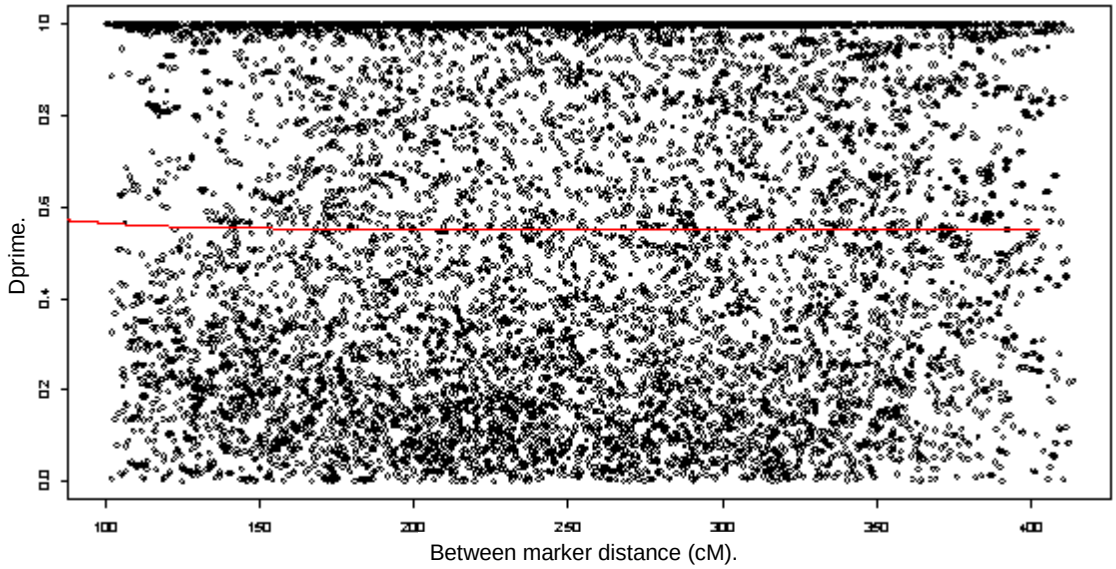


Figure 5-41. Decay of LD (as D') over between marker genetic distance in the original simulated genome before truncation. All possible between chromosome 1 and 2 pair-wise comparisons were made and then re-sampled with replacement to ensure the frequency of D' observations were evenly distributed over distance. The red line is the Malecot curve through the data suggesting LD is elevated but asymptotic at ~ 0.58: the simulated chromosomes are assorting independently.

The distribution of between marker pair distances were compared with the distribution expected under a hypothesis of random placement (Figure 5-42). An excess of low between marker pair distances was observed, this was attributed to the combined effect of drift – which allows allele frequencies to rise by stochastic processes – and the low recombination rate over short genetic distances. Thus if one marker achieves a high MAF it is likely that adjacent markers will do the same.

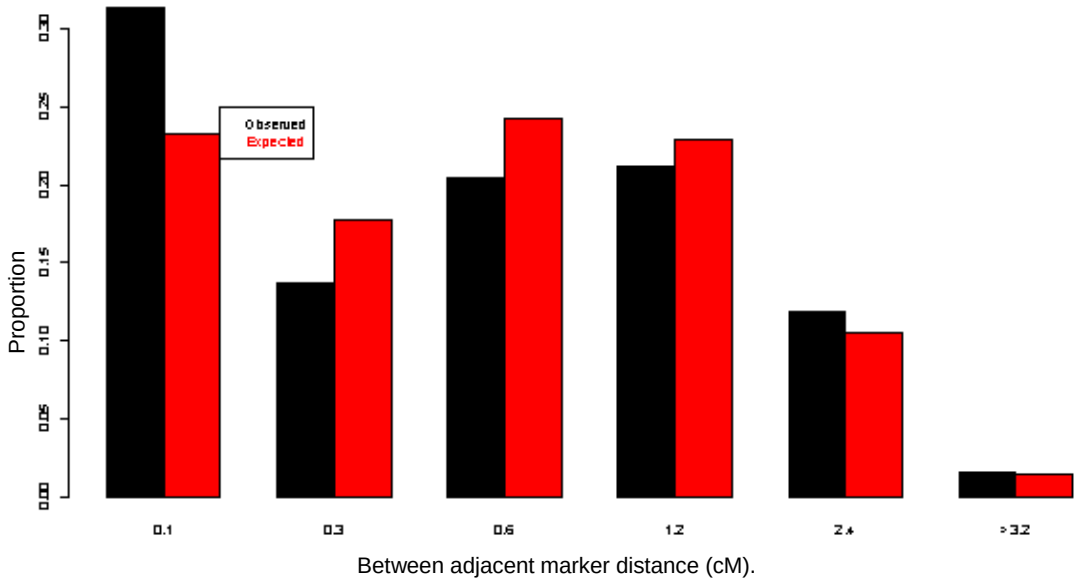


Figure 5-42. Frequency distribution of between adjacent marker distances in the Simulated Cereal dataset compared with the distribution expected if the same number of markers were distributed randomly. [Deviation due to old mutations in close linkage surviving together]

5.3.2 Comparing the simulated cereal with the wheat and barley populations.

The simulated cereal was developed with the intention that it would share some key characteristics with the real populations already obtained. In the context of association mapping the characteristics of interest were LD decay profile and population substructure.

LD decay was conveniently summarised using the coefficients of the Malecot curve for each population (Table 5-14).

Table 5-14. F_{ST} and Coefficients of Malecot curves fitted through the LD decay profiles at different points in the development of the simulated cereal population. Wheat and barley panels included for comparison.

	l	*m*	*t*	F_{ST}
Wheat	0.2333398	0.823812	25.29941	0.30[1]
Barley	0.2664376	0.7166288	11.39081	0.34[2]
Sim A	0.576305	0.9519296	7.118838	0.54[3] (0.29)[4]
Sim Pop 1 (at point A)	0	0.8839606	2.595124	-
Sim Pop 2 (at point A)	0.08477326	0.906748	3.167137	-
Sim Pop 3 (at point A)	0	0.933778	2.636282	-
Sim Pop 4 (at point A)	0	0.936731	2.611797	-
Sim B	0.06406045	0.7153612	39.58621	0.02
Sim C	0.07492169	0.5883465	18.14633	≈0

Table 5-14 shows that the simulated cereal has been developed with broadly similar characteristics to the wheat and barley populations. The age (t) of the simulated population is similar to the age of the barley population, which suggests that the genetic distance over which association between marker and QTL could be detected as association will be similar. The wheat population has a higher value for t suggesting LD decay may be more rapid and the maximum genetic distance over which association will be detectable will be less. The rate of LD decay will therefore influence the marker density necessary to achieve a given statistical power.

Considering population structure; the confounding effects of structure on association mapping will influence both the degree of fixation and amount of phenotypic variation accounted by sub-population structure. Table 5-14 shows that F_{ST} achieved in the simulated cereal is somewhat higher than that seen in the other populations but this may, in part, be attributed to the more sophisticated model underlying the AMOVA. Where the population

[1] Wheat population structure defined in terms of whether varieties had the 1B1R rye translocation or not.
[2] Barley population structure defined in terms of the combination of seasonal habit and row number of the ear.
[3] Simulated cereal population structure defined in terms of the groups and populations illustrated in **Figure 5-32**. Where population division had not occurred the undifferentiated population was sampled as if populations and groups existed.
[4] Bracketed number is F_{ST} calculated in terms of populations only i.e. ignoring the hierarchical group/population structure.

sub-structure of the simulated cereal is defined only in terms of population – ignoring group – the value for F_{ST} obtained is similar to that for wheat and barley.

I conclude that the simulated cereal can be used as a basis for simulating association mapping studies in larger panels, allowing empirical estimates of statistical power to be made.

6 Statistical Power in Association Mapping.

The estimation of the power (1-β) (β ~ the false rejection rate) available in an association mapping experiment is an important step in the analytical process since it informs experimental design, allows comparison between analytical approaches and encourages caution when experimental results appear inconsistent with the power estimated to be available.

Power is routinely estimated either by simulating phenotype from real genetic data, perturbing an existing phenotype or simulating both phenotype and genotype. Association mapping performed using the simulated data allows an estimate of power since true discoveries can be identified and counted following each simulation run. Repeated cycles of simulation and analysis give increasingly precise estimates of power, conditional on the models underlying the simulation.

Here I report the estimation of power in the Wheat and Barley datasets using both simulated and real phenotype data and a range of regression models, I also use simulated genotype data to investigate the effect of increasing panel size and marker numbers.

6.1 Estimation of power.

The principle is to conduct an association study in a data set where the position and characteristics of at least one QTL is known. By observing the success rate for detecting what is known it is possible to infer the power to detect unknown QTL.

In human genetic epidemiology it is common practice to estimate the statistical power of a study in advance. Several online resources exist to facilitate this process (Gordon *et al.* 2002 Purcell *et al.* 2003). However these tend to focus on case control studies and ignore the possibility of quantitative traits. Problems of spurious association due to the effects of uncontrolled population structure are ignored or avoided by modelling TDT studies.

In plant genetics, the formal consideration of power is less common; notable exceptions are Zhao *et al.* (2007) and Yu *et al.*(2006) who both perturb an existing phenotype by the addition of genetic effects from randomly selected SNPs, and then attempt to detect the SNPs chosen by association with the modified phenotype. This approach has the merit of masking the added phenotype with real signals from population structure and environmental variation

and so gives a realistic estimate of the power available to find the SNP if it were part of the genetic architecture of a similar trait. However it seems likely that this approach will give an inflated estimate of power because the criterion used for true discovery is to find the SNP which has been used to adjust or define the phenotype. To mimic a real association study the definition of success should be measured in terms of detecting SNPs which are in association with the un-genotyped causative locus by reason of their physical proximity (linkage); the signal from these may be considerably weaker.

An alternate approach, therefore, is to remove the causative SNP from the genotype prior to association analysis and attempt to locate its position by detecting association with other SNPs. In this case the definition of success is more complex and must be descried in terms of detection of association within a given genetic interval of the causative SNP.

Neither of the above approaches allows for full partitioning of all discoveries into false discoveries and true discoveries[1] and the definition of power is in terms only of finding QTLs added to the phenotype tested. More general questions as to what level of power might be expected with a range of numbers of causative loci, differing heritability, effect sizes, etc can only be made from repeated simulation of traits from genotype data. Power in this latter case might be expressed as 'proportion of simulations in which at least one causative locus was discovered' or 'average proportion of causative loci discovered'. A recent example of this approach is reported by Yu *et al.* (2008) who simulate a phenotype in a simulated maize population and then report the power to detect the ***causative*** SNP; the power to locate by association is not reported.

6.2 Comparison between experiments.

For a fixed α or q-value it is tempting to assume that such estimates of power give an objective basis for comparison between experiments. However this simple approach is misleading since the different regression models and experimental designs vary in terms of the efficiency with which they control spurious association with the result that high power

[1] In this context false discoveries are statistically significant associations not attributable to physical proximity to the causative polymorphism. These will comprise associations significant due to Type III error (Type III error being as defined by Mosteller (1948): 'the error of correctly rejecting the null hypothesis for the wrong reason') and Type I errors. Similarly, true discoveries are statistically significant associations found in close physical proximity to the causative polymorphism but, of course, these may also include some Type I errors.

could reflect either or both of high Type III and low Type II error rate. The key problem, therefore, is to allocate significant associations to the correct alternate hypotheses.

In reality identifying the correct alternate hypothesis for each significant result is impossible. However, it is possible to use completely simulated phenotypes and the objectives of association mapping to generate rules under which associations may be allocated to 'True Discovery' and 'False Discovery'. I have defined the process as:

A. Determine a significance threshold appropriate to the experiment (i.e. specific to the panel, genotype, phenotype and regression model).

B. Specify a genetic interval around the causative QTL within which significant associations will be considered 'True'.

C. Use the interval from (B) to empirically determine power and false discovery rate for each experiment.

6.2.1 Efficiency score.

Since the analysis of each experiment is completely independent but the criteria for discovery are held in common, the resulting empirical estimates of power and FDR may then be used as the basis of an objective assessment of the relative merits of different experimental approaches. There remains the problem of balancing low empirical FDR and high power; both are desirable, and indeed one without the other is of no value whatsoever. To facilitate this evaluation I adopted a simple efficiency score (E) defined in terms of an 'ideal' association study in which all statistically significant associations occur within a stated interval of the causative loci and no causative locus remains undiscovered. Thus complete efficiency occurs when all QTL have been discovered, and no false discoveries made.

$$E = q(1 - f)$$

Where; q = proportion of QTL discovered and f = empirically observed FDR – both conditional on the stated genetic interval.

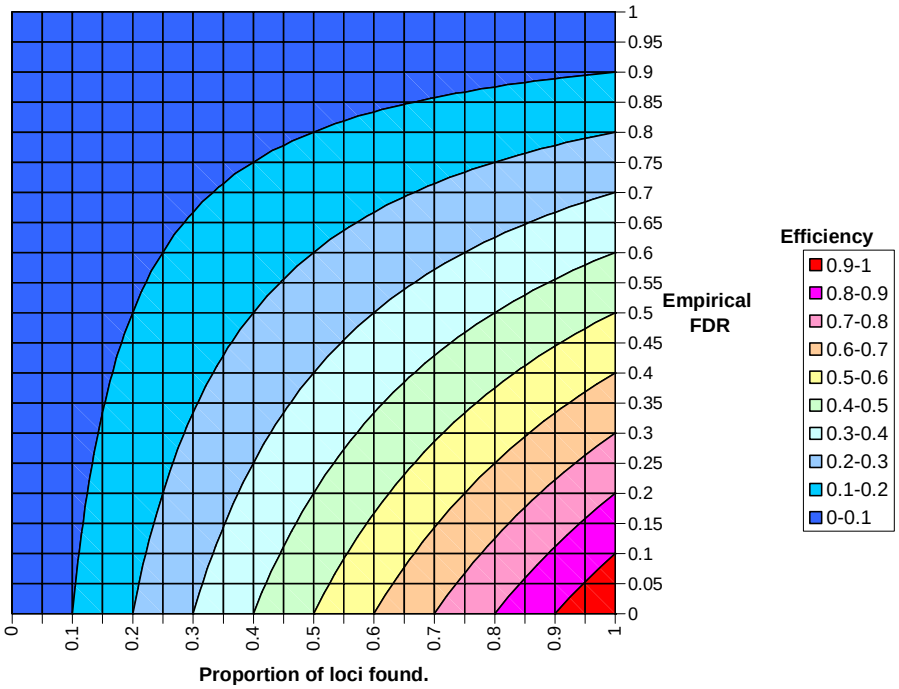


Figure 6-1. Efficiency as a function of both proportion of loci found and the empirical FDR.

6.2.2 The power ratio.

It is also useful to consider the power obtainable under the null. In this case the significant markers found (for whatever reason) would be expected to be randomly distributed in the genome with respect to the QTL, contingent on the spatial distribution of the marker panel used.

If the marker panel were randomly distributed then the equation derived by Bishop *et al.* (1983) could be used to calculate *P*, the probability that a randomly selected marker will fall in a specified window of size *d* cM (i.e. the QTL) in a genome of length *l* cM comprising *c* chromosomes all of length greater than *d.*

$$P = \left[1 - \frac{2c\left(\left[1-x\right]^{n+1} - \left[1-2x\right]^{n+1}\right)}{n+1}\right] - (1-2xc)(1-2x)^{n} \qquad \text{where } x=d/l$$

From this it would be possible to calculate the expected number of discoveries with multiple windows and multiple markers. However, the marker panels under consideration are not distributed randomly (**Figure 5-2** and **Figure 5-15**) and so the power under the null for a

given number of QTL and markers has been determined empirically using Monte Carlo simulation: for *m* significant markers and *q* qtl the panel of mapped markers were randomly sampled (without replacement) to allocate *m* marker locations to represent the statistically significant discoveries and *q* marker locations to represent the position of the causative polymorphism of the QTL. The proportion of *q* 'discovered' when one or more of the markers fell within a specified distance of *q* was considered to be the power under the null. This estimate was averaged over 1000 simulation runs for each of the tested combinations of *q* and *m*.

A useful metric for method to method comparison is then the ratio of the proportion of QTL discovered by that method to the proportion expected to be discovered under the null. This is calculated over multiple simulations of the phenotype (6.3.1.2) and Monte Carlo simulation. I refer to this as the power ratio. The power ratio gives greatest weight to high power achieved with few significant discoveries (see; Figure 6-2).

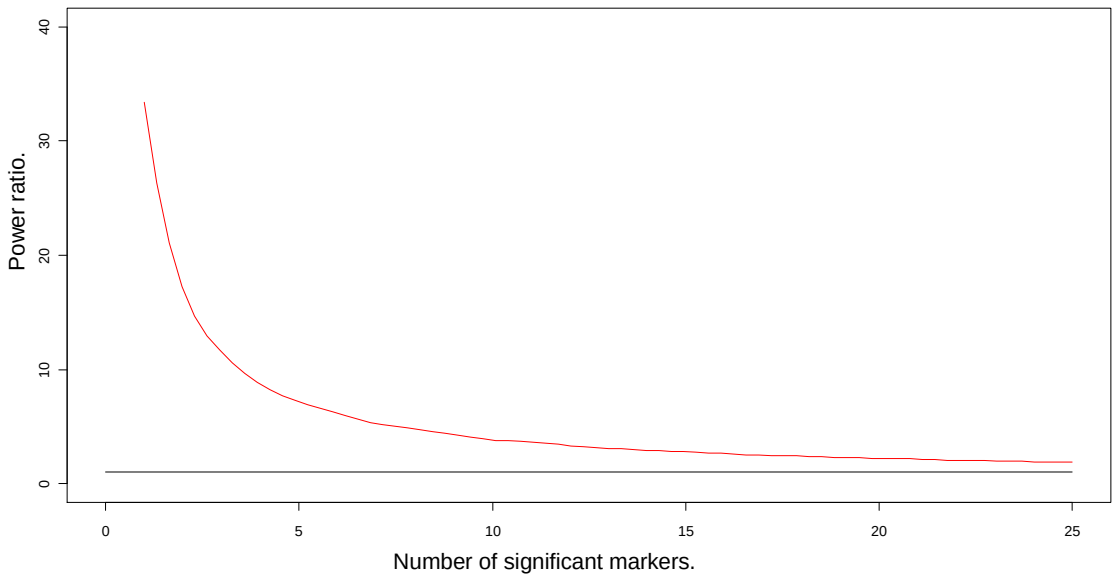


Figure 6-2. The power ratio (red curve) for the detection of two QTL in an experiment with 80% power, over the number of significant (P≤0.05 Bonferroni corrected) markers reported. The power ratio is independent of false discovery rate since it measures only the gain in success over that expected under the null, it emphasises the achievement of obtaining high power with few markers. [Horizontal black line is the expectation under the null. Curve calculated for the marker distribution in the barley dataset.]

Power ratio varies with the significance threshold since a lax threshold allows more discoveries which in turn increases the likelihood of discoveries being made under the null.

With a severe threshold discoveries are rare and the differential between the null and the observation is greater.

6.3 Simulated genotypes and phenotypes.

The wheat and barley datasets have been previously described (Chapter 5) and offer the possibility to predict, by simulation of phenotype, the statistical power of the association mapping experiments with real data. Moving beyond simulations based in real phenotype data, fully simulated genotype data (**5.3**) offer the possibility of investigating the effects of increased panel size and increased number of markers.

6.3.1 Simulated phenotypes

6.3.1.1 Perturbed Phenotypes.

Following the method of Yu *et al*.(2006), existing phenotypes (HFN, Sp Wt, Protein and Yield in Wheat and Yield in Barley) represented as BLUEs of variety means derived from trial data (**5.2.1.1**) were perturbed by the addition of a genetic effect from each available marker in turn and then conducting a new association analysis. The effect size of the added QTL was varied (0, 0.2, 0.4, 0.6, 0.8, 1.0 times the phenotypic standard deviation) to give 6*m* (*m* = No. of markers) perturbed phenotypes per simulation run. For clarity; genotype *1* had the effect added to its phenotype, genotype *0* had no effect added. Under these conditions, for a given allele frequency, the proportion of the phenotypic variation explained (π) at each effect size is given by:

$$\pi = \frac{p(1-p)k^2}{(p(1-p)k^2 + 1 - \frac{1}{n})} \qquad \textit{where n is large} \qquad \approx \left[1 + \left(\frac{1}{p(1-p)k^2}\right)\right]^{-1}$$

Where *n* is the sample size, *p* is the minor allele frequency of the marker (assuming a binary genotype) and *k* gives the multiple of the phenotypic standard deviation added, i.e. the effect size.

Power was calculated as an average over all markers and the proportion of variation explained was calculated as an average over all markers for a given effect size. Power was defined both as; (i) the proportion of occasions when the causative polymorphism was found to be significantly associated with the perturbed phenotype and (separately) (ii) the

proportion of occasions where association with the perturbed phenotype was detected within 10cM of the known location of the causative marker (causative marker having been excluded from the analysis).

6.3.1.2 Fully simulated phenotypes

Phenotypes were simulated by selecting, at random, n_l loci to control the simulated trait. The maximum genetic component of the simulated trait was 9 units which occurred when the binary haploid genotype comprised the '0' allele at all n_l causative loci. The effect size for each locus was $9/n_l$ for genotype '0' and zero for genotype '1'.

To add an effect due to population structure the simulated barley phenotype was adjusted according to whether individuals fell into winter, spring, 6-row or 2-row groups. Specifically the added effects were; Sp/2row=+5, Sp/6row =+7, W/2row=+0, W/6row=+2. Individuals with phenotype unknown for these traits were assigned a trait combination randomly, the assignment was made in proportion to the frequency of the trait combination observed in the data. Thus the assignment is unbiased but will reduce power. The vector of sub-population structural effects on the simulated wheat phenotype for the *n* individuals (**S**) was determined from the Q matrix (*k*=7) and an arbitrarily assigned effect for each sub-population and is given by:

S=QK

Where **Q** is the *k* by *n* Q matrix from STRUCTURE and *k* is a *k* by 1 vector comprising the numbers 1 to *k.*

In the case of the simulated cereal the effect of population structure was added on the basis of the first two principal coordinates (PC1<0 and PC2<0 = 0, PC1<0 and PC 2 >0 = 2, PC1>0 and PC2<0 = 4 and PC1>0 and PC2<0 = 6), this approximated to the population allocations of the individuals since the four populations are distributed into the four graphical quadrats in the first two principal coordinates.

The variance of this combined genetic and structurally derived phenotype was determined over the variety panel and used as an estimate of genetic variance (V_G). Environmental variance (V_E) was added as ~N(0, V_E) to achieve heritabilities (h^2) of 0.5 and 0.9 where:

$$h^2 = \frac{V_G}{(V_E + V_G)}$$

The process of creating a simulated phenotype is illustrated graphically in Figure 6-3 to Figure 6-6.

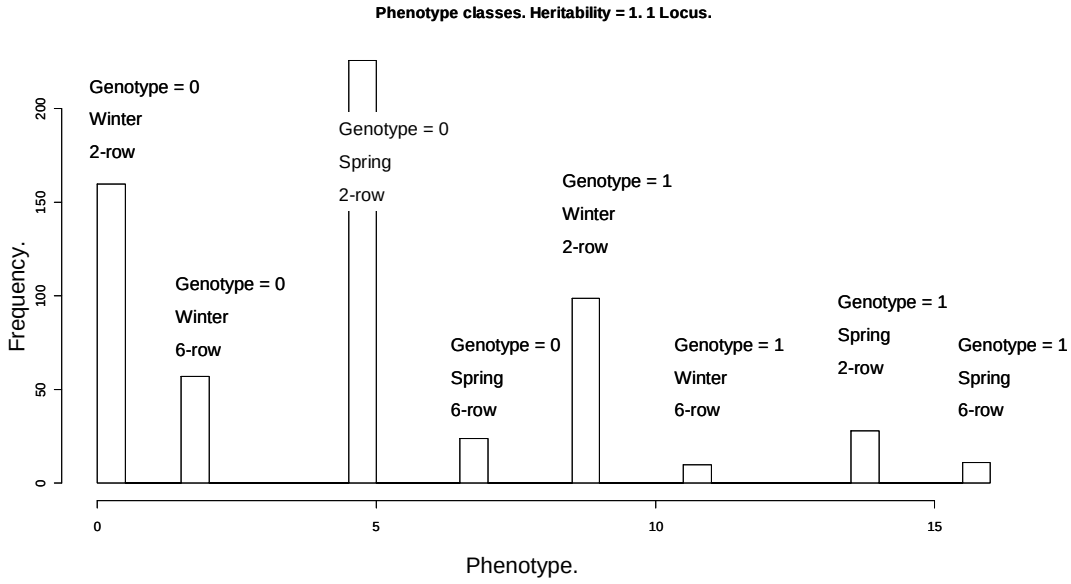


Figure 6-3 Frequency distribution of phenotypic states illustrating the combined effect of genotype and population structure. Phenotype simulated from one locus and $h^2 = 1$.

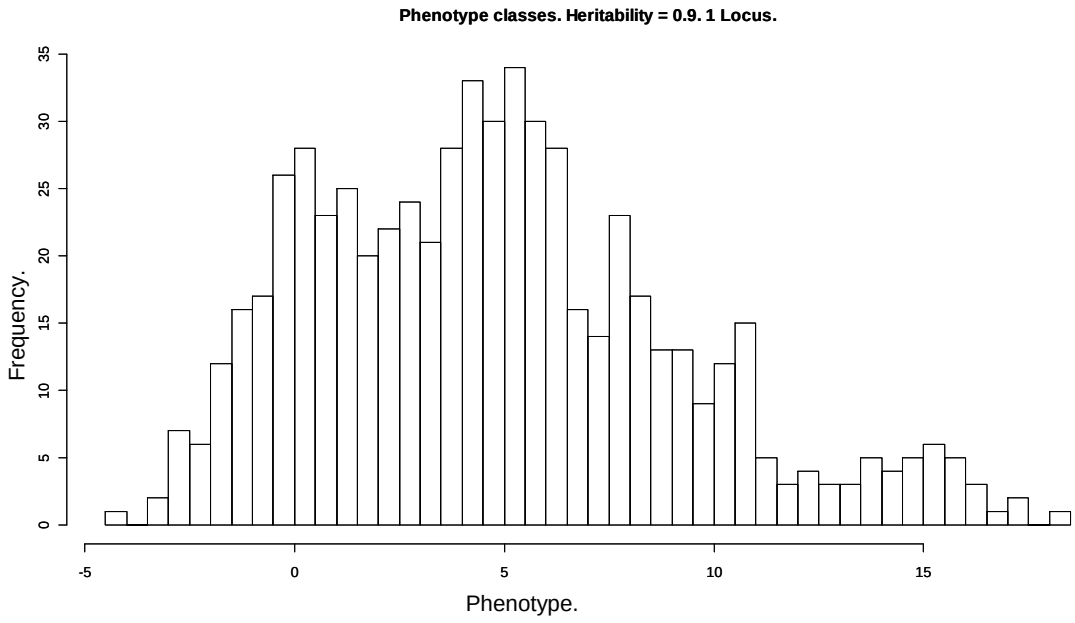


Figure 6-4 Frequency distribution of phenotypic states illustrating the combined effect of genotype, population structure and heritabiliity. Phenotype simulated from one locus (as previous figure) and $h^2 = 1$ (red bars) or $h^2 = 0.9$ (clear bars).

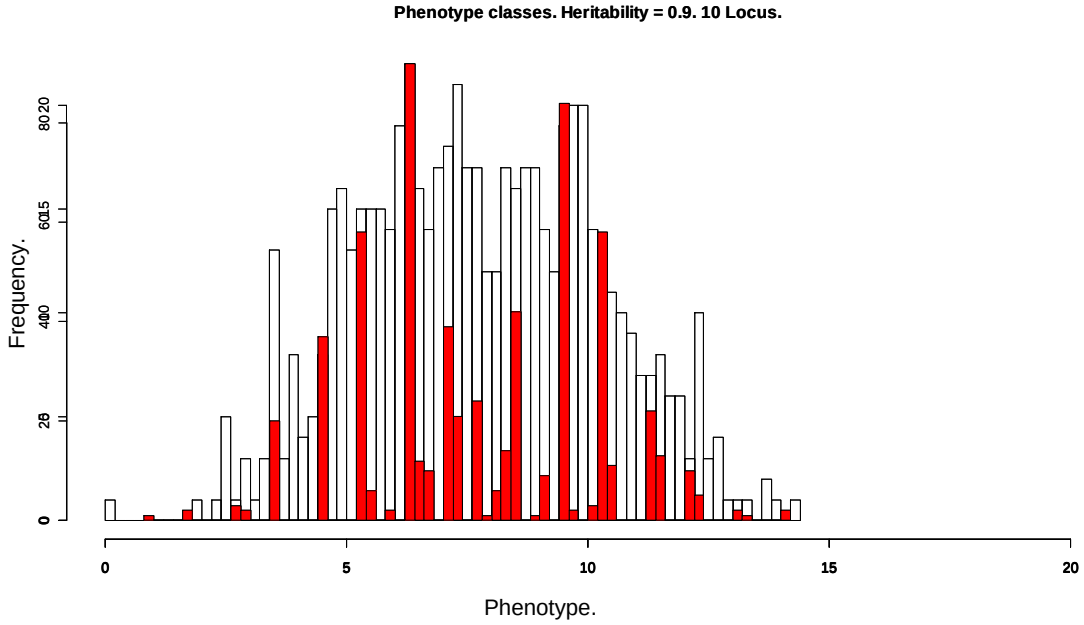


Figure 6-5 Frequency distribution of phenotypic states illustrating the combined effect of genotype, population structure and heritabiliity. Phenotype simulated from 10 loci and $h^2 = 1$ (red bars) or $h^2 = 0.9$ (clear bars).

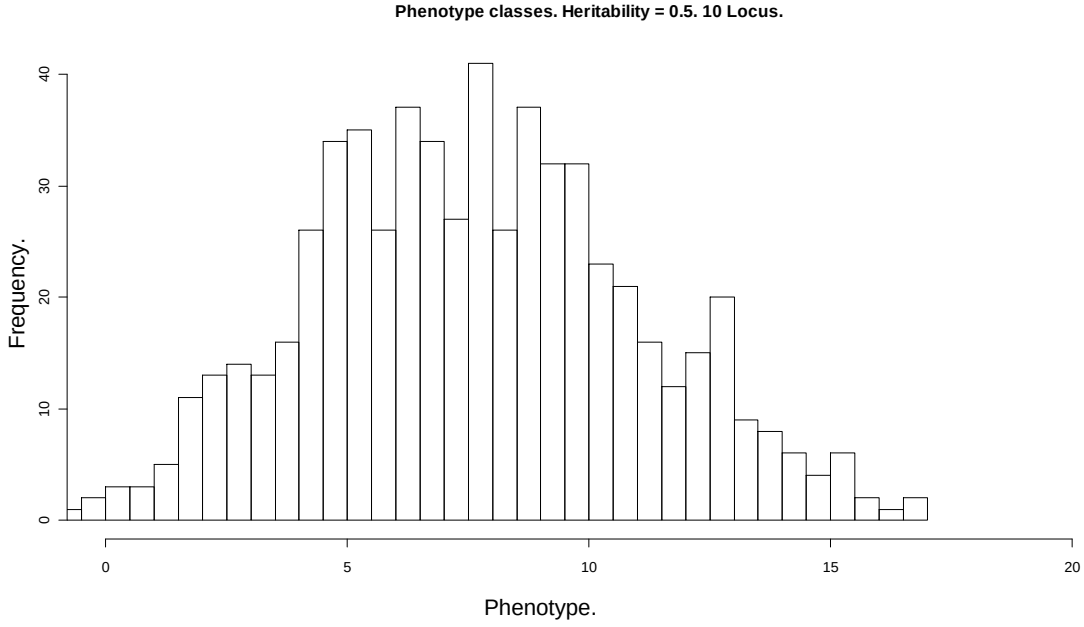


Figure 6-6 Frequency distribution of phenotypic states illustrating the combined effect of genotype, population structure and heritabiliity. Phenotype simulated from 10 loci (as previous figure) and $h^2 = 0.5$.

6.3.2 Minor Allele Frequency (MAF) in Simulations.

Phenotypic simulations were conducted with MAF set to 0.1.

6.4 Results

6.4.1 Perturbation of existing phenotype.

6.4.1.1 Wheat

Power varies with effect size, regression model and phenotype.

Figure 6-7A shows that in the case of the unperturbed HFN phenotype there is genome wide inflation of the test statistic using all regression models except MM. Genomic control can be used to remove the generalised inflation and
Figure 6-7B shows that this treatment has produced probability distributions which lie mainly under the null. Some observed excess significance at the extreme of the distribution suggests statistically significant association with the trait and this excess is the source of any statistical power the test may have.

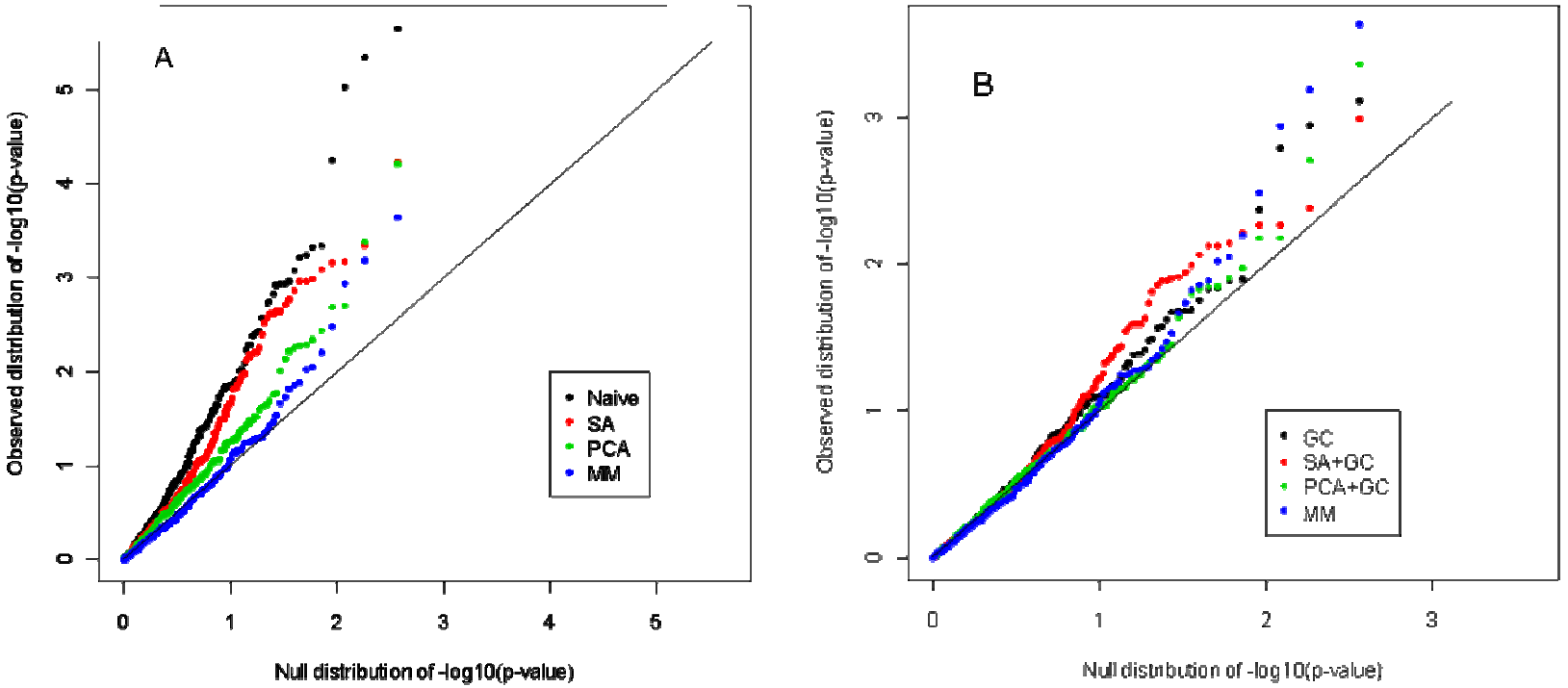


Figure 6-7 Association with the unperturbed HFN phenotype. Inflation of the estimated statistical significance relative to the null (y=x) is evident over all markers in Naïve association. Although SA and PCA reduce the inflation somewhat only the MM returns the majority of distribution to that expected under the null (A). Using GC alone or in conjunction with SA and PCA (B) controls the inflation. The MM attributes greater significance to the four most associated markers than the other methods – indicative of potentially higher power (B).

Considering the power to detect perturbations to the phenotype and considering first the regression model, with reference to Figure 6-8 to Figure 6-11; both structured association (SA) and correction with PCA (eig) approaches retain most power, however Figure 6-7 A shows that these methods must produce an excess of significant results (false positives) attributable to the incomplete control of spurious association due to population structure and kinship. This is evidenced by the generalised deviation from y=x for these methods. If genomic control is used to control this then the power to detect falls substantially.

Naïve association offers power approaching 100% (data not presented for wheat but see for example, figure Figure 6-20 (A) below), but this simply reflects genome-wide spurious association giving very high levels of false discovery. Applying genomic control reveals that the discrimination between true and false discoveries in the naïve model is poor; power is less than 10% for yield and protein and is also lowest of all methods for these phenotypes. In the case of Hagberg falling number and specific weight (Figure 6-9 and Figure 6-10) GC appears to have similar power to SA+GC and PCA+GC. This is an interesting outcome suggesting that inflation of the test statistic is not being controlled by SA or PCA for these phenotypes. This observation is not entirely unexpected since population structure will only lead to spurious association if subpopulations differ with respect to their average value for the phenotype in question. There is, however, evidence that the test statistic was inflated; the distribution of the association results were inflated relative to the expectation under the null.

The mixed model (MM) appears to offer the most consistent performance; it does not exhibit inflation relative to the null (
Figure 6-7), in terms of power it consistently matches or outperforms the other regression models (once they have had GC applied) (Figure 6-8 to Figure 6-11). A particularly interesting effect has already been noted in the case of specific weight and Hagberg falling number, where SA+GC and PCA+GC perform little better than GC but the MM offers power similar to SA or PCA uncorrected. This suggests that spurious association with these traits results from kinship which is largely ignored by SA and PCA but is specifically accounted in the mixed model. This is reflected in the % error accounted (R^2) in the prediction of phenotype from the Q matrix or the first 20 eigenvectors of *k* (Table 6-1) where the both specific weight and Hagberg falling number are least well predicted; indicating that there is little effect on phenotype.

Table 6-1 Prediction of phenotype from models of population structure.

Phenotype	% Error accounted by Q (k=7)	% Error accounted by first 20 eigenvectors of K
HFN	15	32
Protein	42	62
Yield	56	68
Specific Weight	13	34

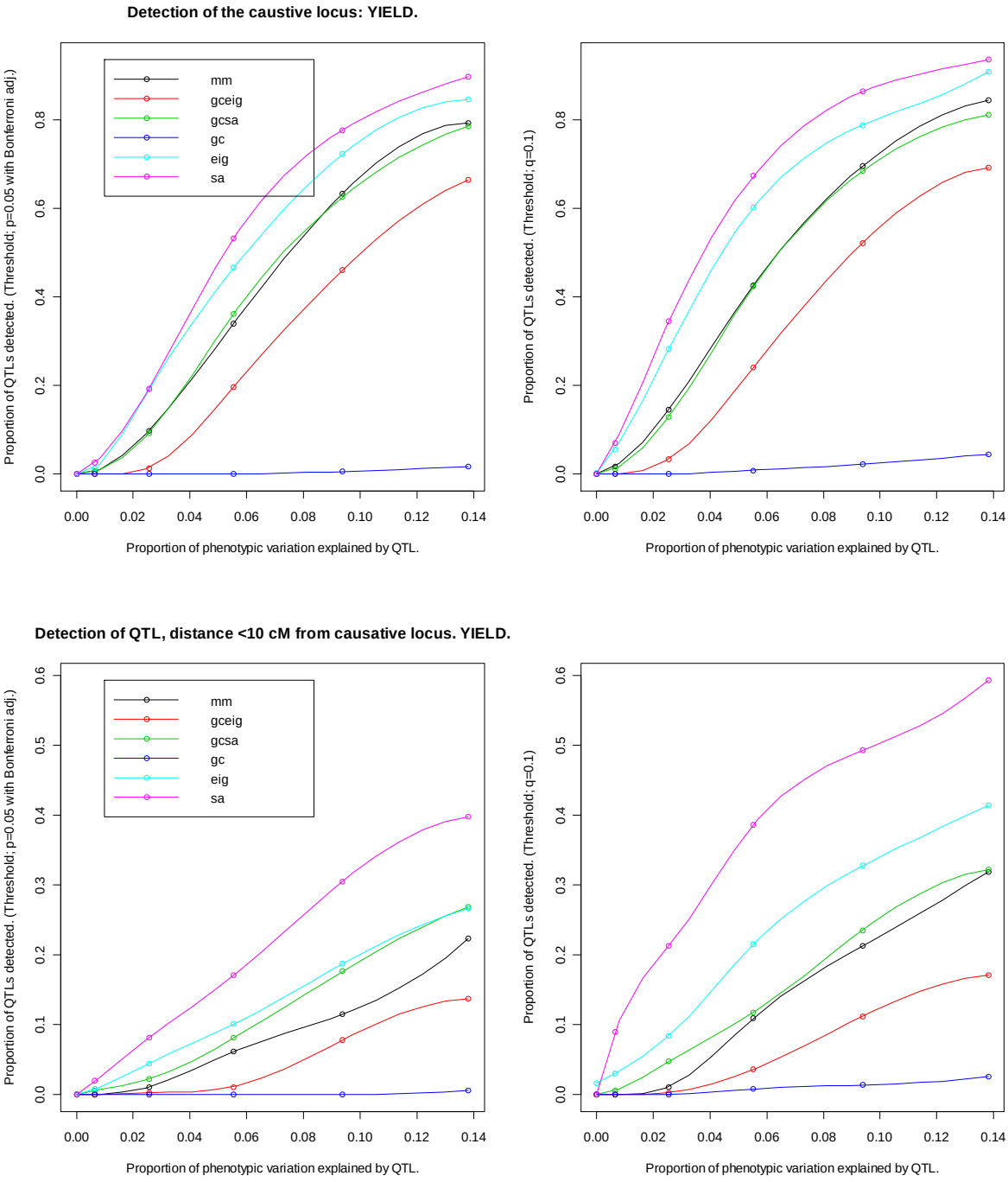


Figure 6-8 Power to detect association with a perturbation to the yield phenotype of wheat. Right hand graphs represent power with significance set at q=0.1, left hand graphs significance is p=0.05 (Bonferroni corrected). Upper graphs represent power to detect the causative polymorphism, lower graphs show power to detect the QTL within a genetic distance of 10cM.

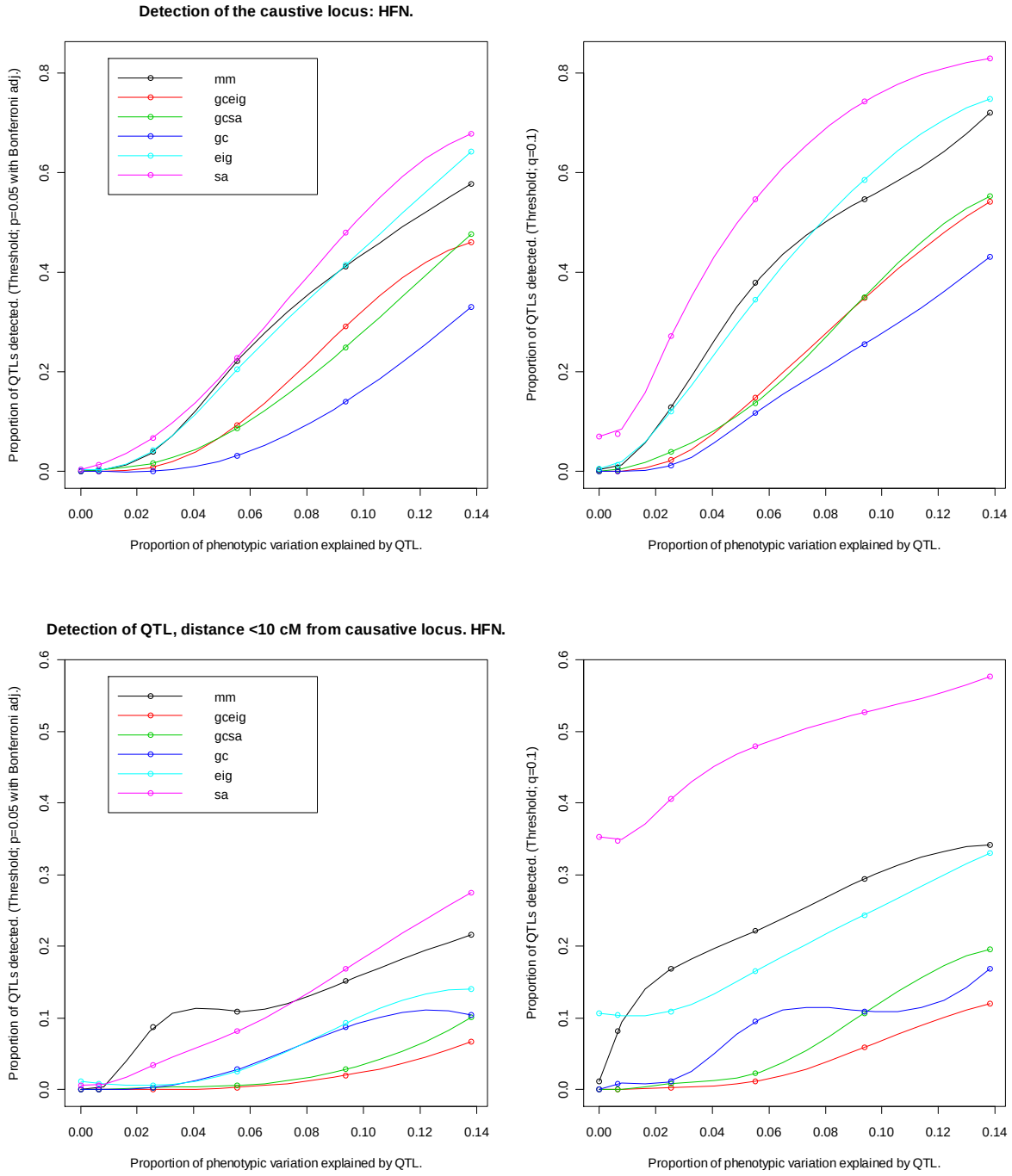


Figure 6-9 Power to detect association with a perturbation to the Hagberg falling number phenotype of wheat. Right hand graphs represent power with significance set at q=0.1, left hand graphs significance is p=0.05 (Bonferroni corrected). Upper graphs represent power to detect the causative polymorphism, lower graphs show power to detect the QTL within a genetic distance of 10cM.

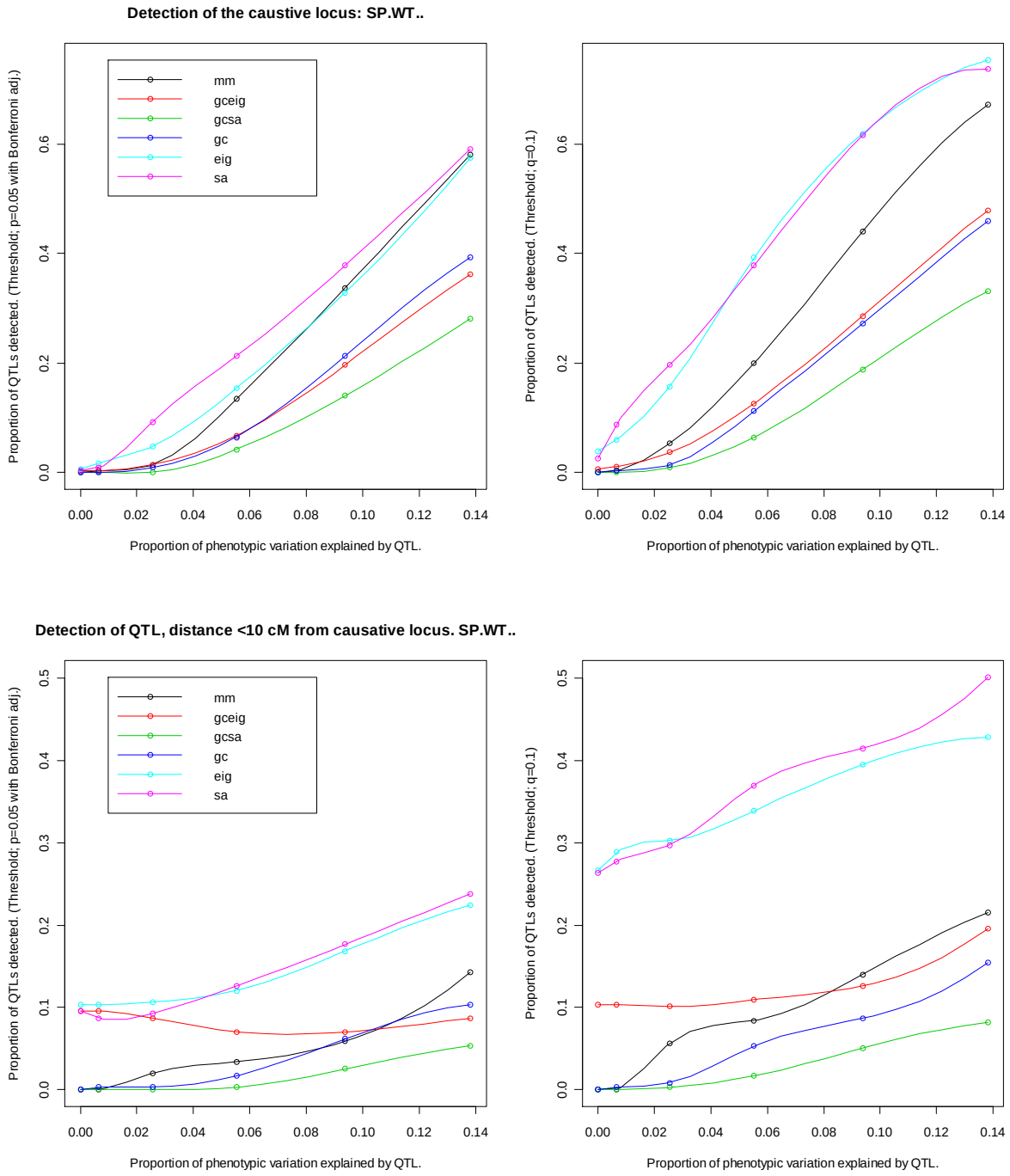


Figure 6-10 Power to detect association with a perturbation to the specific weight phenotype of wheat. Right hand graphs represent power with significance set at q=0.1, left hand graphs significance is p=0.05 (Bonferroni corrected). Upper graphs represent power to detect the causative polymorphism, lower graphs show power to detect the QTL within a genetic distance of 10cM.

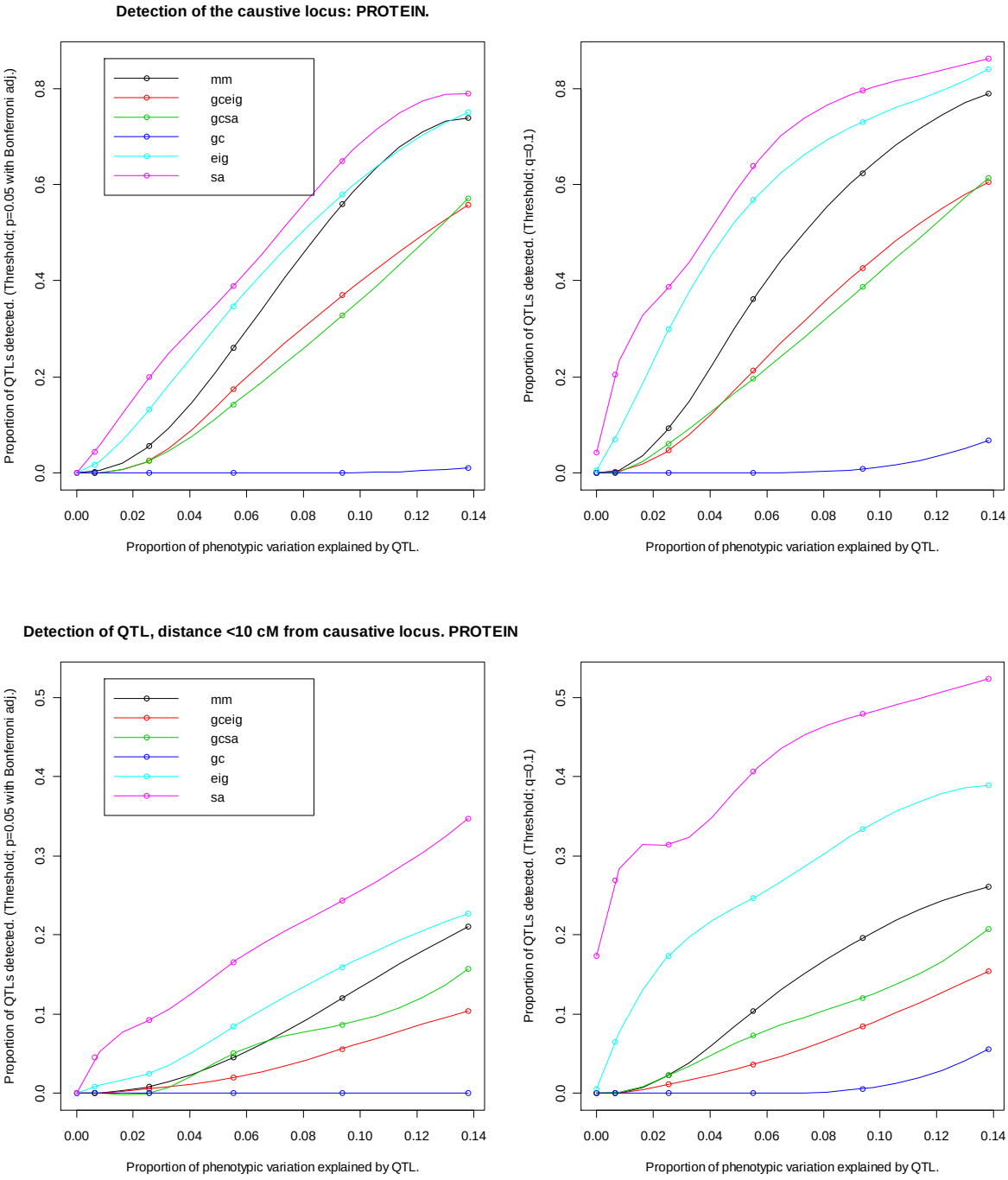


Figure 6-11 Power to detect association with a perturbation to the grain protein phenotype of wheat. Right hand graphs represent power with significance set at q=0.1, left hand graphs significance is p=0.05 (Bonferroni corrected). Upper graphs represent power to detect the causative polymorphism, lower graphs show power to detect the QTL within a genetic distance of 10cM.

6.4.1.2 Barley.

I chose to perturb just the yield phenotype since this is a quantitative trait comparable with the traits examined in wheat.

Figure 6-12 shows that in the case of the unperturbed yield phenotype in the entire dataset (winter and spring barley combined) there is genome wide inflation of the test statistic using naïve association. The MM with GC almost completely controls inflation. Conventional GC controls the inflation but eradicates the signal. It is important to note that the use of winter and spring barley together in this way will have imposed a very strong signature from structure since the average yield of winter barley is about 16% higher than that for spring barley.

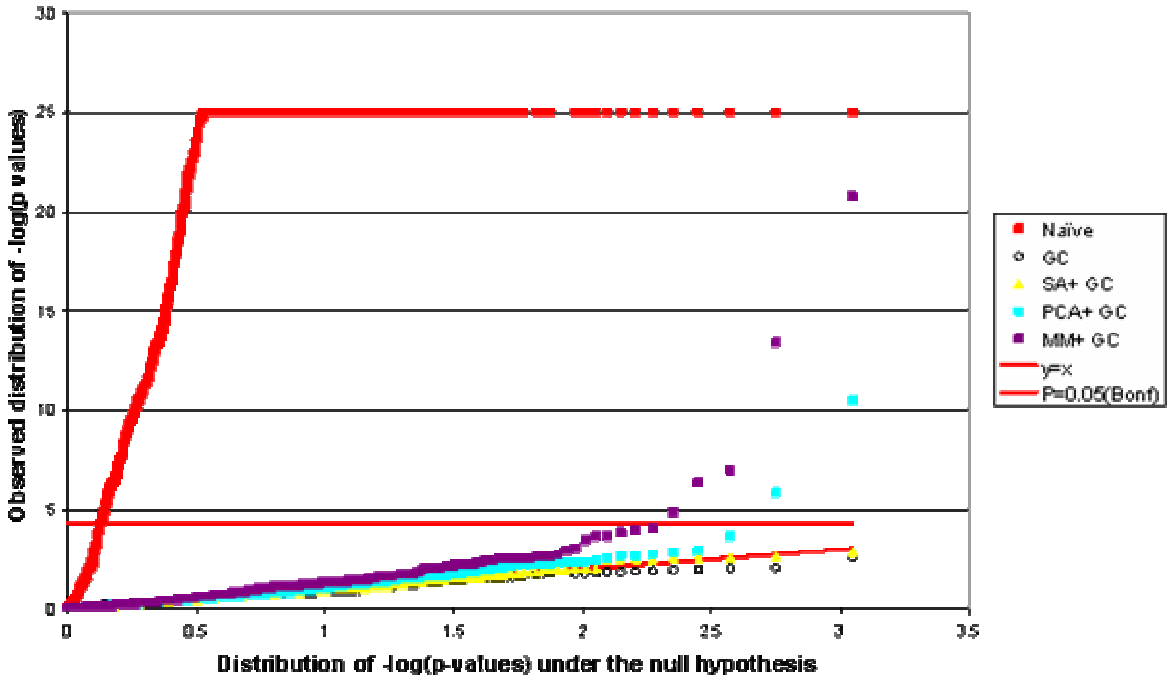


Figure 6-12. PP Plots to show the distribution of $-\log_{10}$(p-values) for association with yield measured in the combined winter and spring barley dataset. The horizontal red line represents a Bonferroni corrected significance threshold of p=0.05. The black line is the expected distribution under the null. Strong signatures from structure are predicted because winter barley out-yields spring barley by 1 t ha^{-1}. Note: -log10 (pvalues) >50 have been set to 50.

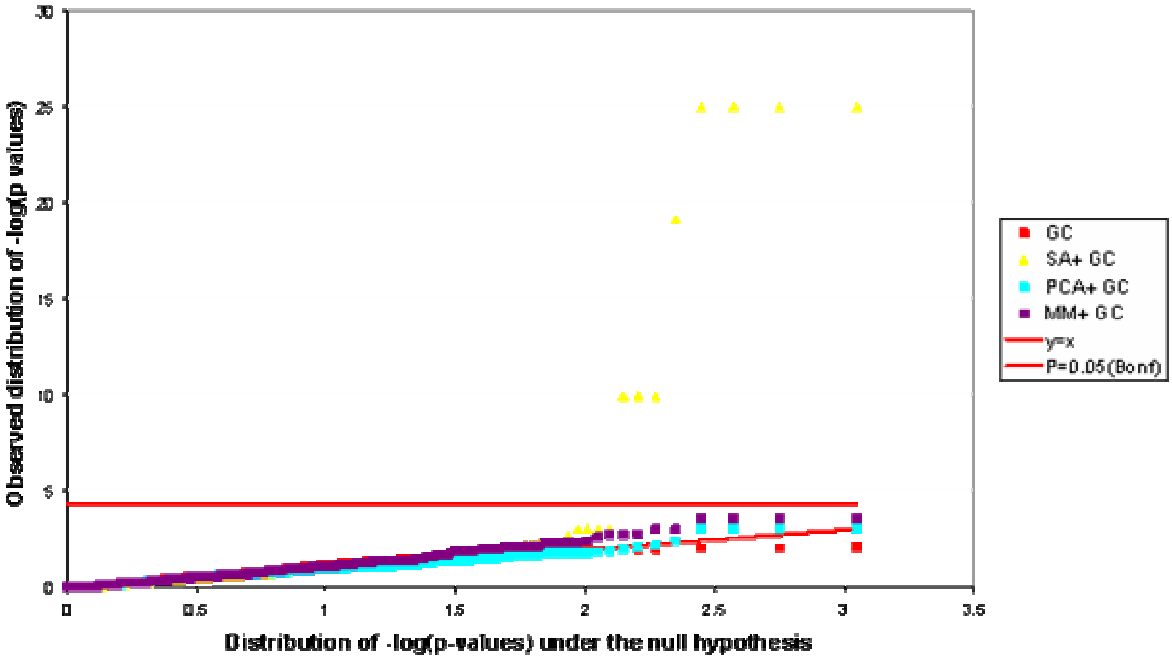


Figure 6-13. PP Plots to show the distribution of $-\log_{10}$(p-values) for association with yield measured in the winter barley dataset. The horizontal red line represents a Bonferroni corrected significance threshold of p=0.05. The black line is the expected distribution under the null. Note: -log10 (pvalues) >25 have been set to 25.

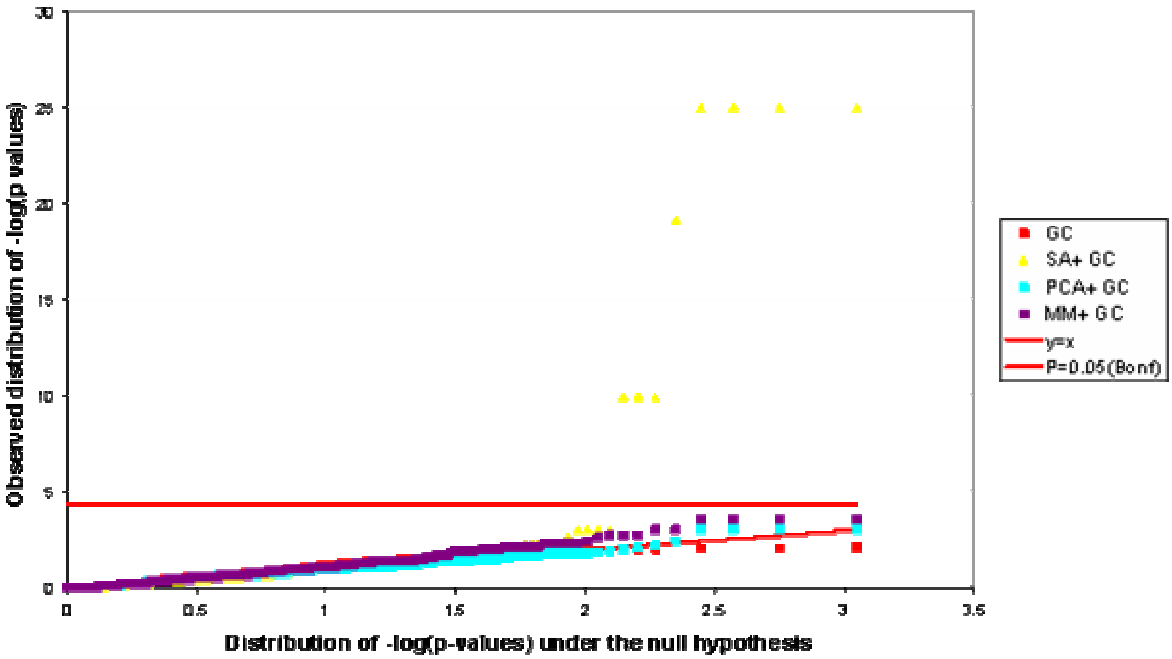


Figure 6-14. PP Plots to show the distribution of $-\log_{10}$(p-values) for association with yield measured in the spring barley dataset. The horizontal red line represents a Bonferroni corrected significance threshold of p=0.05. The black line is the expected distribution under the null. Note: -log10 (pvalues) >25 have been set to 25.

Figure 6-13 and Figure 6-14 show that partitioning the data into winter and spring reduced, but does not irradicate, the inflation due to structure.

Considering the power to detect markers used to perturb the yield phenotype, Figure 6-15 and Figure 6-16 show power within spring and winter datasets respectively. For a locus accounting for 10% of phenotypic variation the power is 45% in spring barley and 65% in winter barley. However, when the two groups of samples are combined (Figure 6-17) the power increases to 80%.

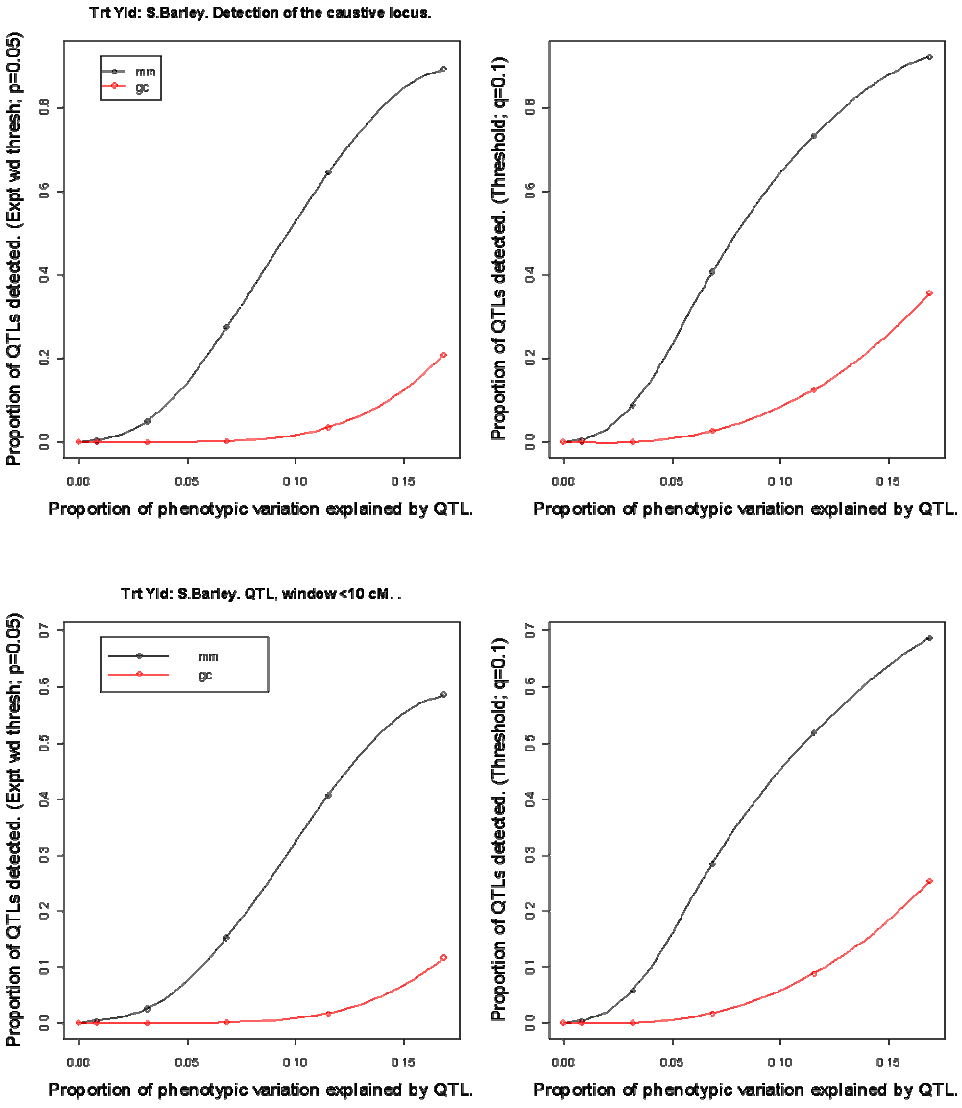


Figure 6-15. Power to detect association with a perturbation to the yield phenotype of spring barley. Right hand graphs represent power with significance set at q=0.1, left hand graphs significance is p=0.05 (Bonferroni corrected). Upper graphs represent power to detect the causative polymorphism, lower graphs show power to detect the QTL within a genetic distance of 10cM.

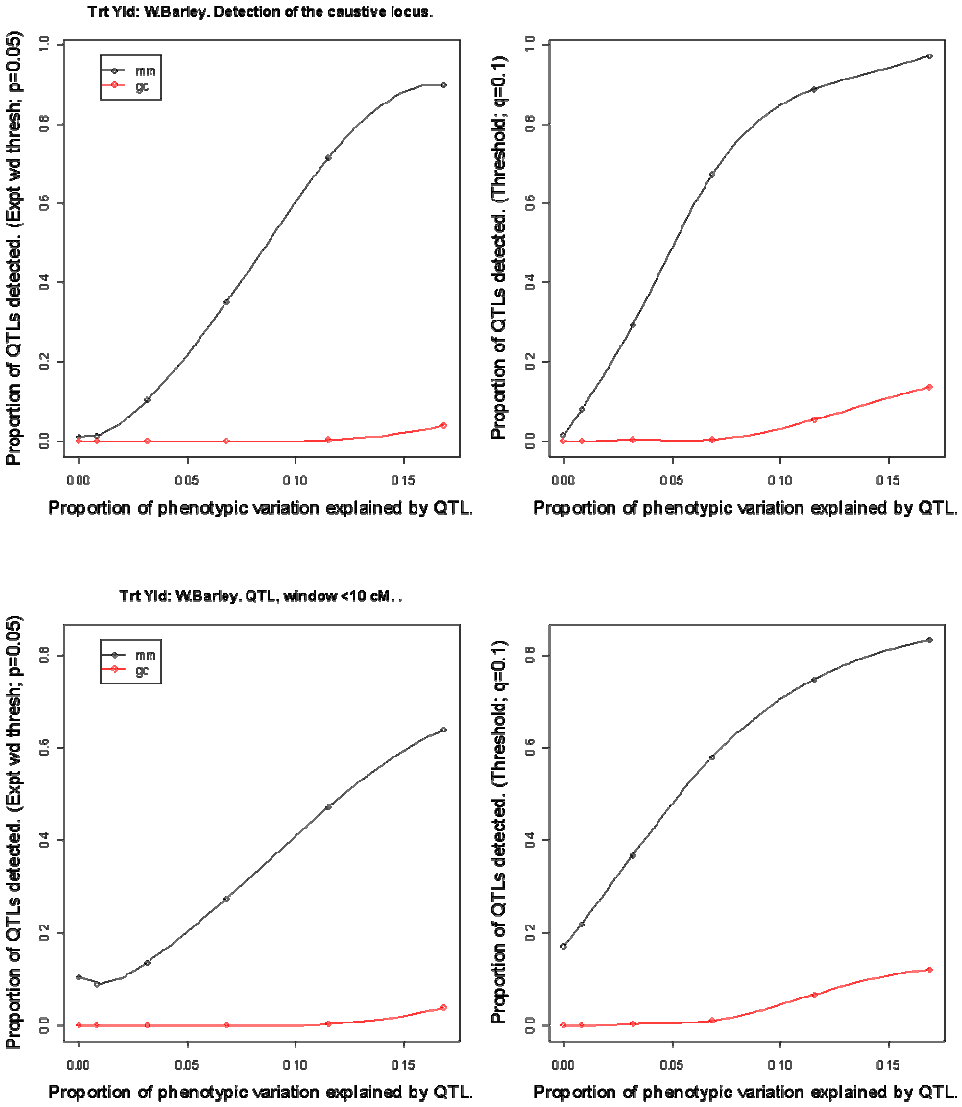


Figure 6-16. Power to detect association with a perturbation to the yield phenotype of winter barley. Right hand graphs represent power with significance set at q=0.1, left hand graphs significance is p=0.05 (Bonferroni corrected). Upper graphs represent power to detect the causative polymorphism, lower graphs show power to detect the QTL within a genetic distance of 10cM.

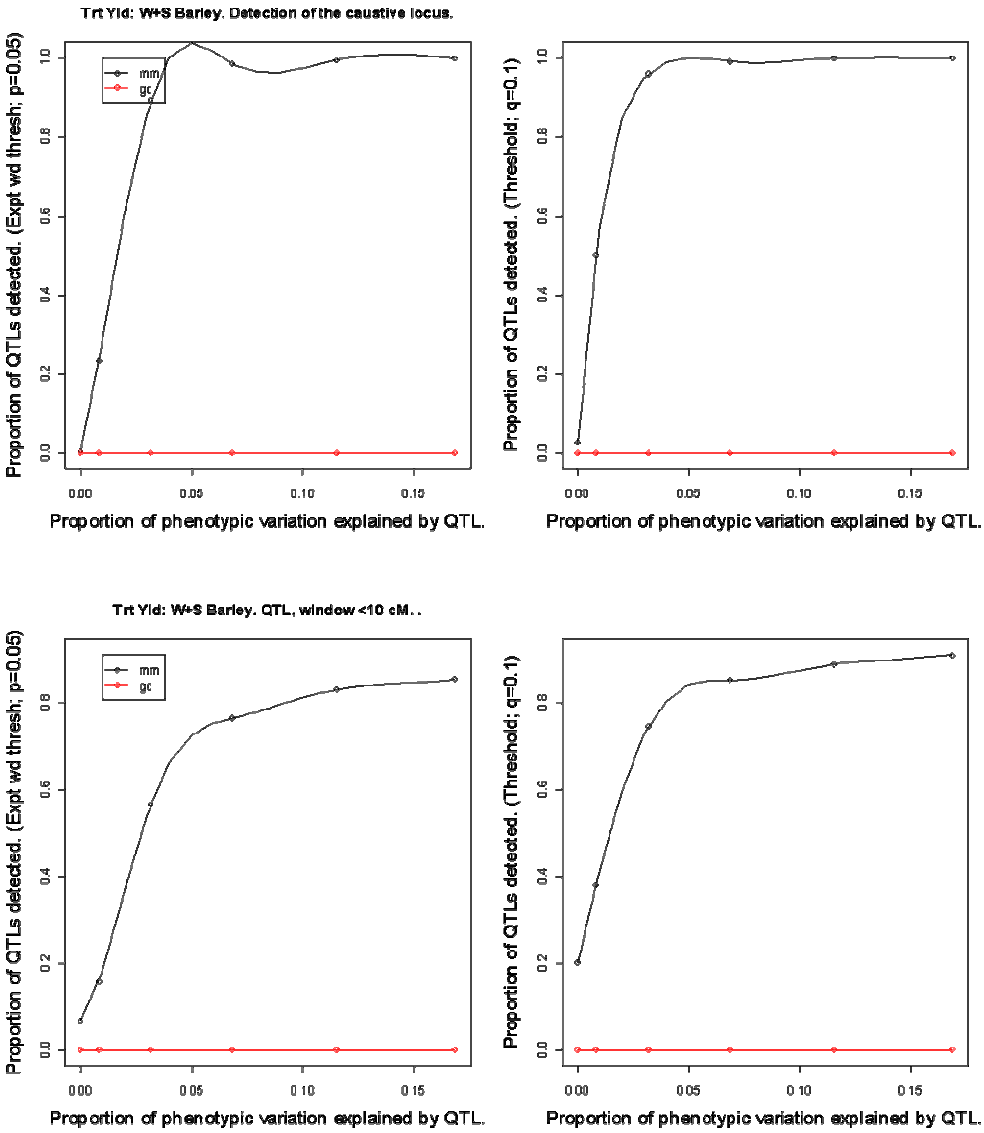


Figure 6-17. Power to detect association with a perturbation to the yield phenotype of winter and spring barley combined. Right hand graphs represent power with significance set at q=0.1, left hand graphs significance is p=0.05 (Bonferroni corrected). Upper graphs represent power to detect the causative polymorphism, lower graphs show power to detect the QTL within a genetic distance of 10cM.

6.4.2 Simulation of phenotypes in wheat, barley and the simulated cereal.

Since the genetic architecture of the simulated phenotype was known the statistical power and false discovery rates could be fully examined. Power was defined as 'The average proportion of causative loci who's location was found over all simulations'. The power ratio (6.2.2) was also calculated. In the estimation of power a true discovery was defined in terms of an interval around the known position of the causative SNP. In addition to power, the proportion of statistically significant discoveries which fall outside the interval was noted and used to calculate the observed false discovery rate (FDR) - which should not be confused with the *q*-values used to determine the significance threshold.

To allow objective comparison between methods the efficiency score (see; 6.2.1) was also estimated. Figure 6-18 and Figure 6-19 show that in wheat and barley SA and PCA offer most power, however as a measure of the best method this is misleading since they also exhibit the highest empirical FDR. The power ratio for SA and PCA shows that these methods derive some of their power from 'guesswork' essentially by making more discoveries they discover more QTL. Applying GC to these methods has the expected result of controlling FDR and shows the power, efficiency and FDR of SA+GC, PCA+GC and the MM to be similar with the MM possibly best overall. The benefit of the MM becomes clearer when the power ratio is considered; here the MM out performs all other methods in both barley and wheat. In essence the MM is achieving comparable power with fewer discoveries; it is a more discriminating method. Arguably the significance threshold for the MM could be lowered, this would increase power without necessarily raising the observed FDR. In these figures the power ratio shows that the gain from regression over random placement becomes modest when the number of loci controlling the trait rises (i.e. the effect size per locus falls) and at very low effect size the regression models perform no better than the null; the experiment has run out of power.

Comparing the estimates of power from fully simulated phenotypes with that estimated from perturbation, it is evident that for QTL which explain similar amounts of phenotypic variation the power observed is also similar. Specifically, the mixed model offers about 20-30% power to detect the perturbation to the phenotype in the quantitative wheat phenotypes when the proportion of variation explained by the QTL is 0.14, this corresponds well with the power

observed when conducting an association study on a simulated phenotype accounting for a similar proportion of variation and in the same panel and genotypes (~20%).

The estimation of power from fully simulated phenotype in the barley genotypes (Figure 6-19) appear to be conservative relative to that estimated by perturbation of the yield phenotype (60% power with ~15% of phenotypic variance explained by the QTL c.f. 80% by perturbation) although given the artificial nature of the simulated trait, and the three-fold difference in sizes between the wheat and barley genomes, this difference is not too concerning.

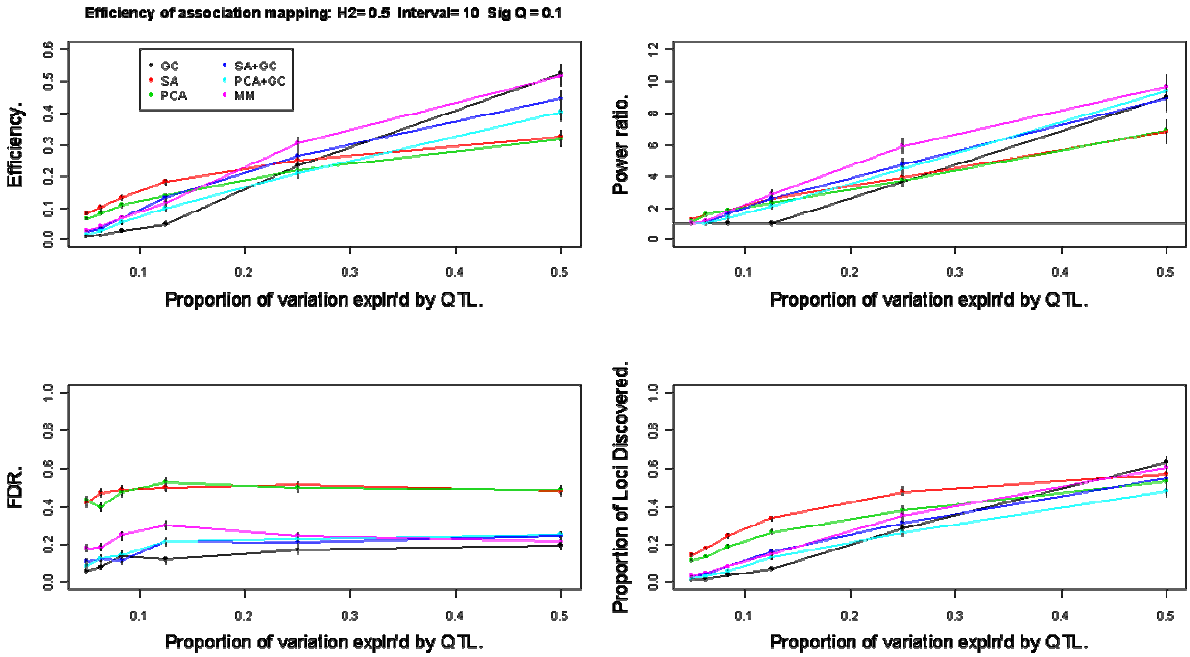


Figure 6-18. Proportion of loci discovered, FDR, Power ratito and Efficiency of regression models used to detect association with phenotype simulated in wheat data over the estimated average proportion of variance explained by each locus. Significance threshold was q=0.1. Detection is considered true within 10cM of the causative locus. Points are average of 400 simulations (error bars ± 1 s.d.).

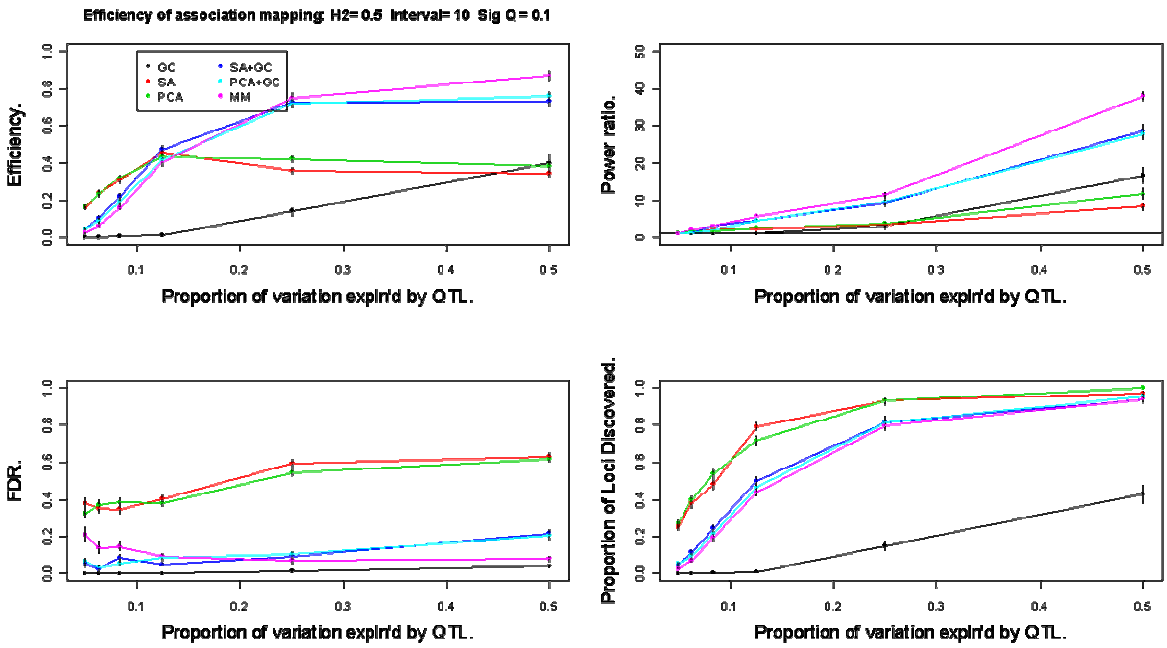


Figure 6-19. Proportion of loci discovered, FDR, Power ratito and Efficiency of regression models used to detect association with phenotype simulated in barley data over the proportion of phenotypic variation explained by the locus. Significance threshold was q=0.1. Detection is considered true within 10cM of the causative locus. Points are average of 100 simulations (error bars ± 1 s.d.).

The simulated cereal gives similar results to both wheat and barley although the data are presented slightly differently (complete figures are presented in Appendix 2, selected figures below). Here the effect of varying both marker and variety panel sizes are considered but for comparison note that when the simulated cereal has panel sizes 150 individuals and 500 markers and loci account for 12.5% of phenotypic variance (Figure 6-25 - similar to the wheat panel) the power statistics are similar in magnitude to those observed in wheat (Figure 6-18). Similarly when the panel sizes are 600 individuals and 1000 markers the performance is similar to barley Figure 6-23 c.f. Figure 6-19 (similar to the barley panel).

Using these simulated genotype data to extend beyond the variety and marker panels available experimentally: Association analysis of the simulated cereal reveals similar features to those seen in wheat and barley; inflation of the test statistic (Figure 6-20, Figure 6-21), greater efficiency with the mixed model, and a similar average proportion of loci discovered for a given panel size as that seen in using the perturbed phenotypes in wheat and barley. (Whilst considering Figure 6-20 B note the effect of GC on uncorrected association is to

reveal a distortion in the distribution of p-values which essentially leaves no signal of association; later figures should be interpreted with this in mind).

As the panel of individuals and markers is increased there is an increase in power. The combined effect of these variables is presented as surface plots (**Figure 6-26** to Figure 6-29). For loci accounting 50% or more of phenotypic variance power for even the smallest panels is greater than 80% (**Figure 6-22**). For more modest effects the power surfaces summarise the relationship between power and the two dimensions of panel size.

Looking ahead to Chapters 7 and 8 a locus effect size of about 10% of phenotypic variance is typical, Figure 6-27 shows that for 12.5% the predicted power with panels of the order of 1500 markers and 1200 varieties is ~80%. At 10% variation accounted the power falls to 70-80% in a similar panel size (**Figure 6-28**). These values are all calculated with a discovery window of 10cM.

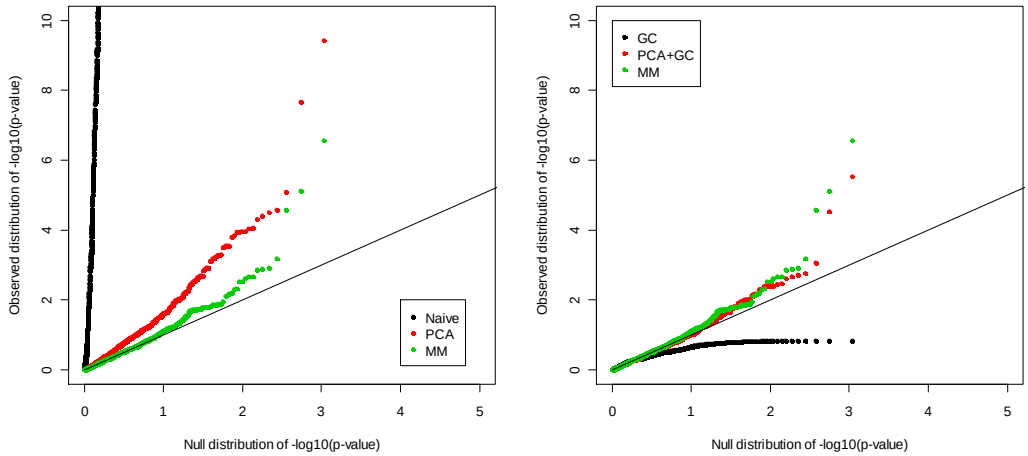


Figure 6-20 Association with a simulated phenotype in the simulated cereal (h^2=0.5) controlled by ten loci. Usng naïve association, extreme inflation of the estimated statistical significance is evident over all markers. Although PCA reduces the inflation somewhat only the MM returns the majority of distribution to that expected under the null (A). Using GC alone or in conjunction with PCA (B) controls the inflation. The MM attributes greater significance to the most associated markers than the other methods (B) – indicative of potentially higher power once the false positive rate has been controlled.

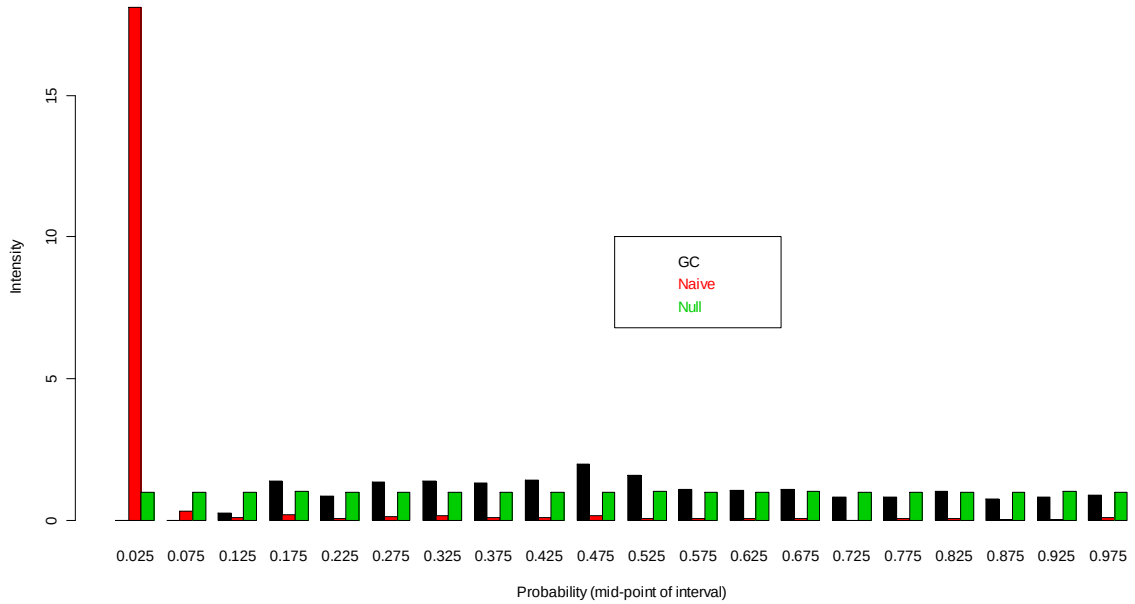


Figure 6-21 The origin of the apparent over-correction to naïve association in the previous figure: Although GC returns a distribution with the mean probability expected under the null this is achieved by indiscriminate experiment-wise reduction in significance, the resulting deficit of low p-values causes the PP plot to lie below the diagonal. This erasure of significance results in a significant loss of power.

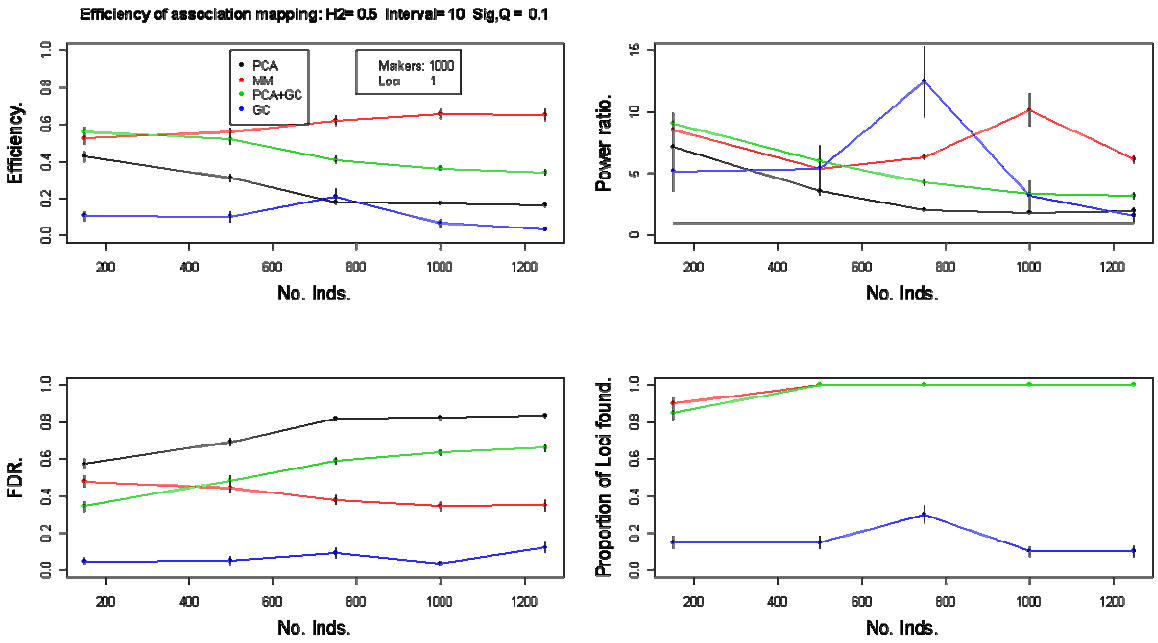


Figure 6-22. Power, FDR and Efficiency of regression models used to detect association with phenotype simulated in the simulated cereal data over number of randomly selected individuals in the panel. Loci account for 50% of phenotypic variation. Genotypes consisted of 1000 randomly selected markers. Significance threshold was q=0.1. Detection is considered true within 10cM of the causative locus. Points are average of 100 simulations (error bars ± 1 s.d.).

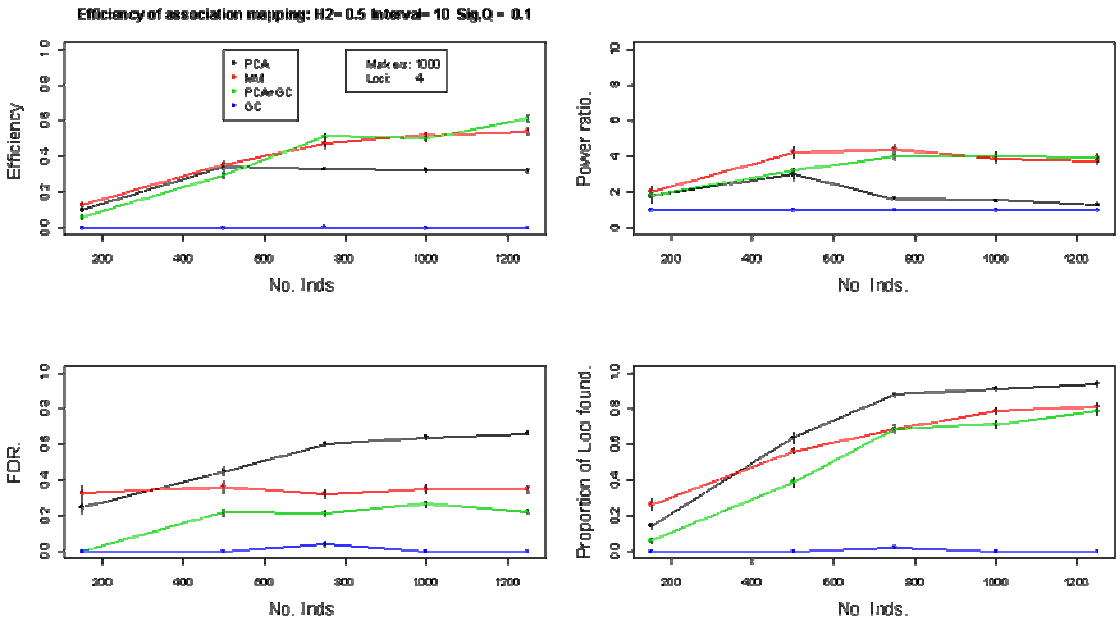


Figure 6-23. Power, FDR and Efficiency of regression models used to detect association with phenotype simulated in the simulated cereal data over number of randomly selected individuals in the panel. Loci account for 12.5% of phenotypic variation. Genotypes consisted of 1000 randomly selected markers. Significance threshold was Q=0.1. Detection is considered true within 10cM of the causative locus. Points are average of 100 simulations (error bars ± 1 s.d.).

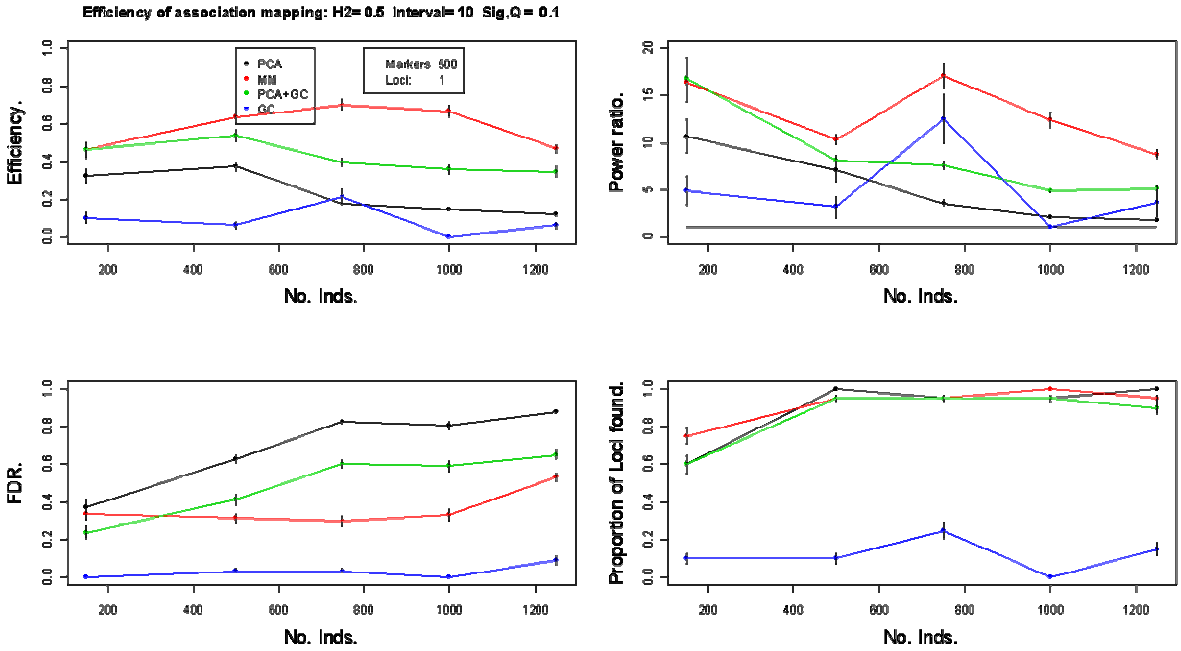


Figure 6-24. Power, FDR and Efficiency of regression models used to detect association with phenotype simulated in the simulated cereal data over number of randomly selected individuals in the panel. Loci account for 50% of phenotypic variation. Genotypes consisted of 500 randomly selected markers. Significance threshold was q=0.1. Detection is considered true within 10cM of the causative locus. Points are average of 100 simulations (error bars ± 1 s.d.).

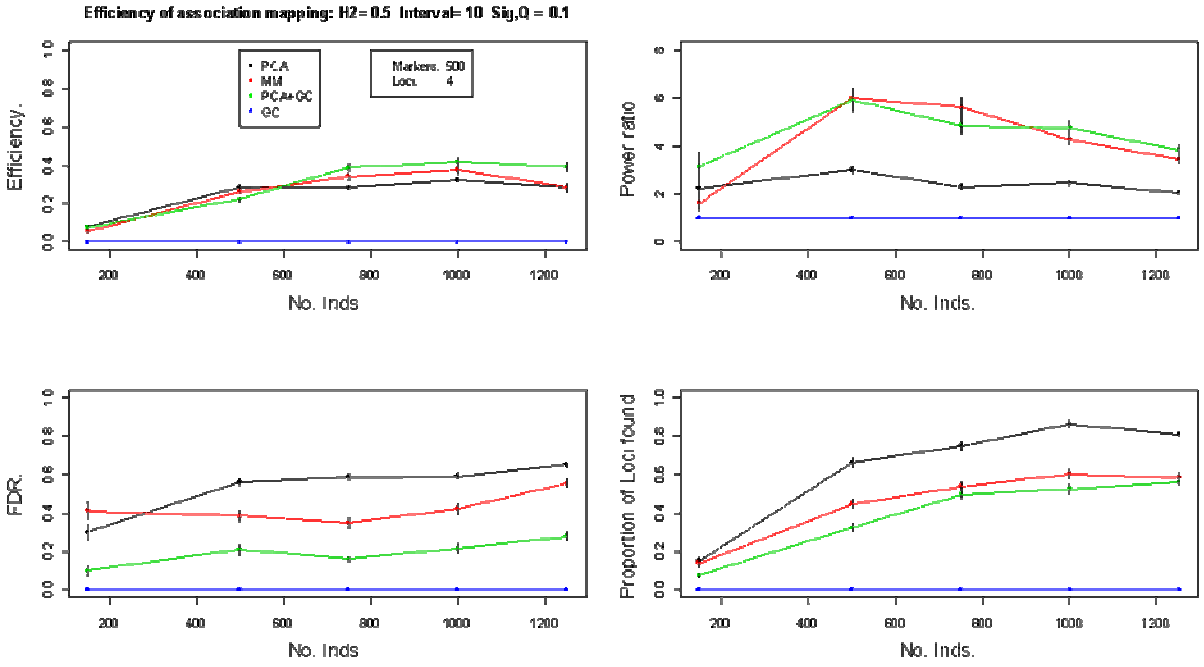


Figure 6-25. Power, FDR and Efficiency of regression models used to detect association with phenotype simulated in the simulated cereal data over number of randomly selected individuals in the panel. Loci account for 12.5% of phenotypic variation. Genotypes consisted of 500 randomly selected markers. Significance threshold was q=0.1. Detection is considered true within 10cM of the causative locus. Points are average of 100 simulations (error bars ± 1 s.d.).

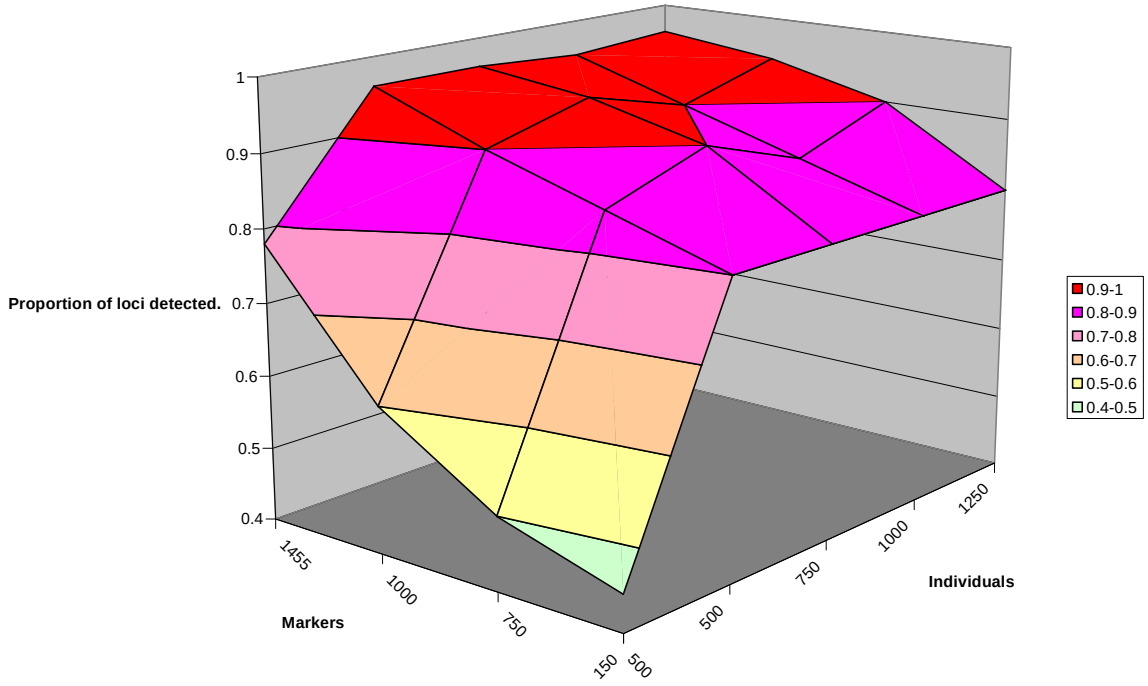


Figure 6-26. The proportion of loci discovered increases with both marker number and number of individuals. Example is of simulated phenotype in the simulated cereal, loci account for 22.5% of phenotypic variation. Based on MM analysis. Window for discovery =10cM.

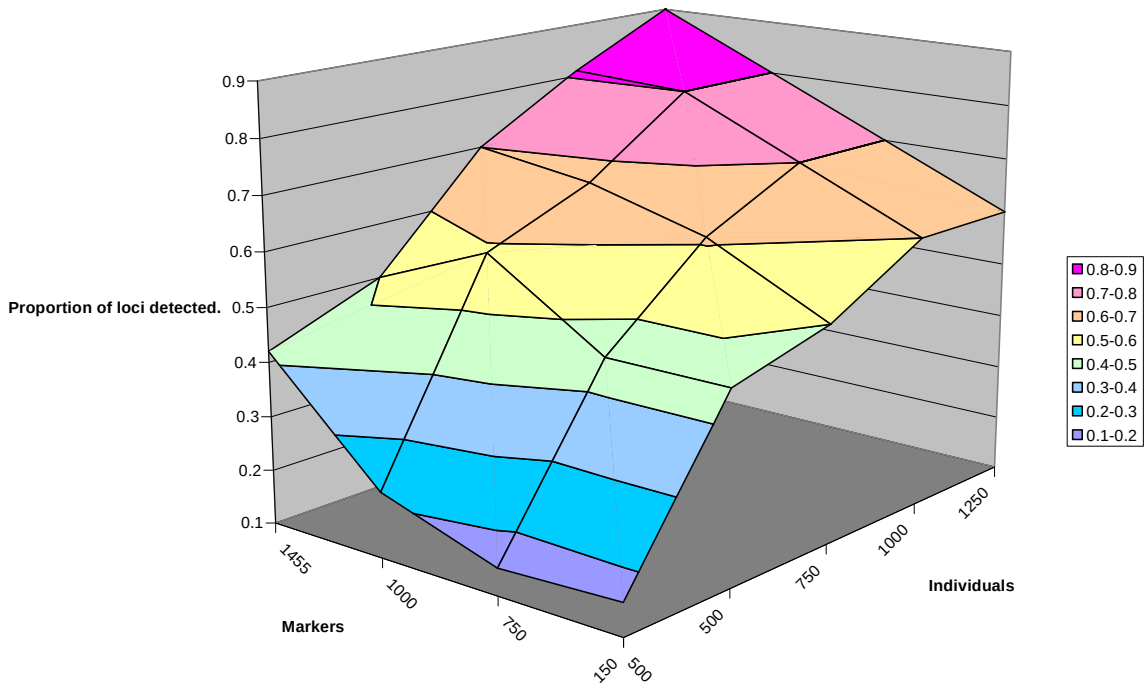


Figure 6-27. The proportion of loci discovered increases with both marker number and number of individuals. Example is of simulated phenotype in the simulated cereal, loci account for 12.5% of phenotypic variation. Based on MM analysis. Window for discovery =10cM.

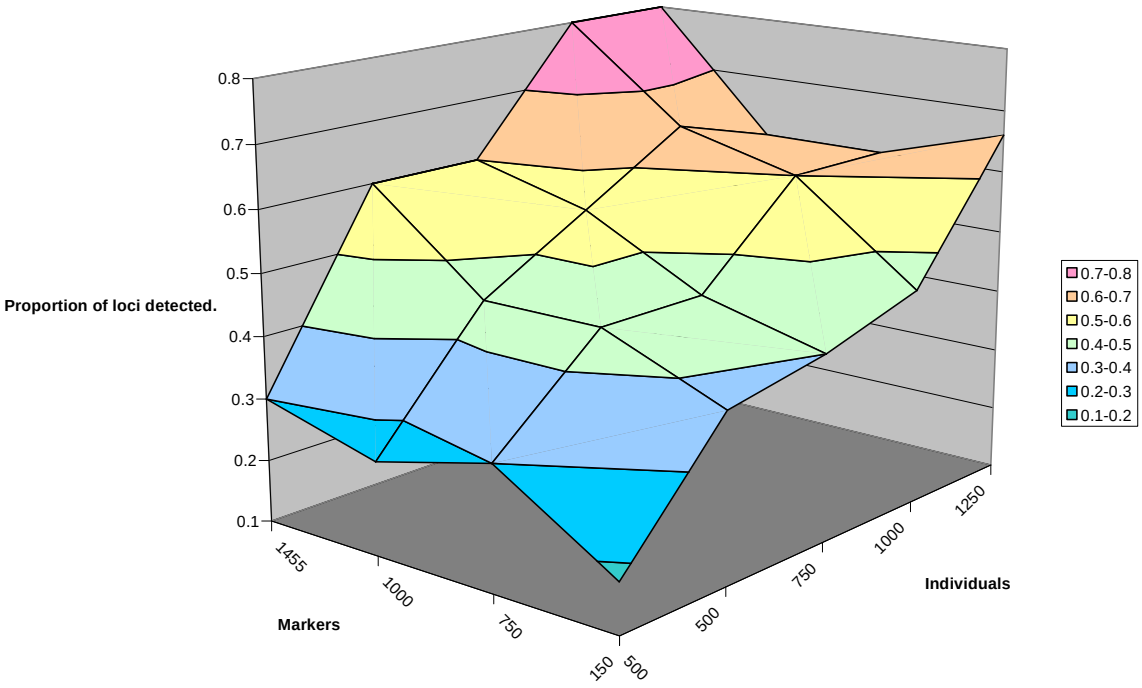


Figure 6-28. The proportion of loci discovered increases with both marker number and number of individuals. Example is of simulated phenotype in the simulated cereal, loci account for 9% of phenotypic variation. Based on MM analysis. Window for discovery =10cM.

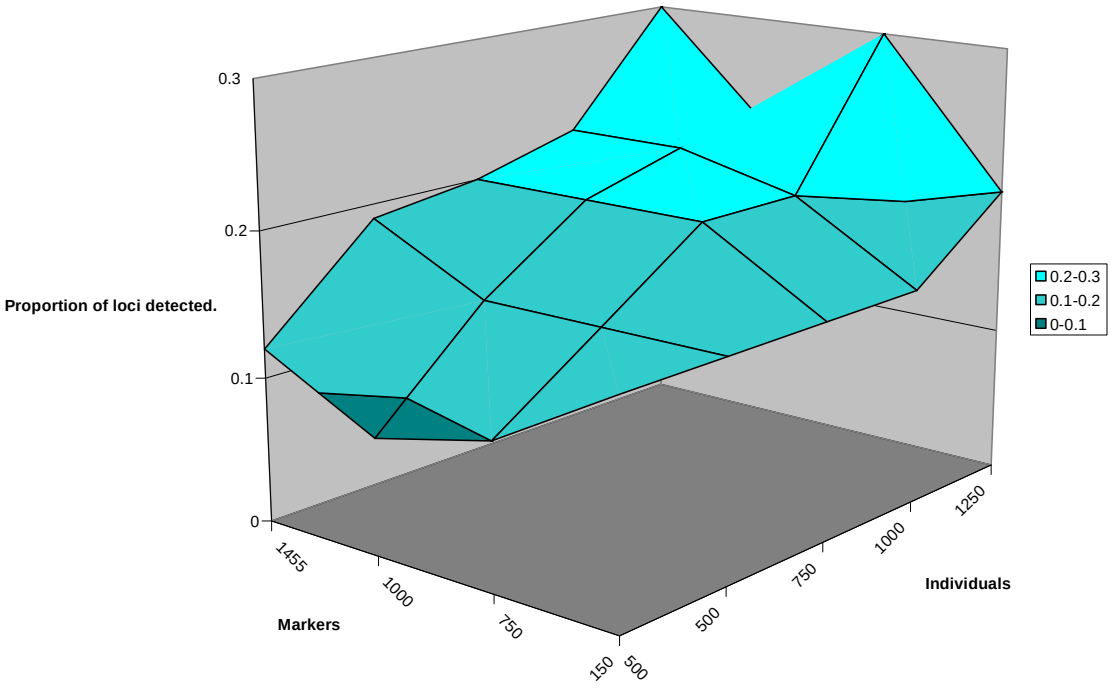


Figure 6-29. The proportion of loci discovered increases with both marker number and number of individuals. Example is of simulated phenotype in the simulated cereal, loci account for 5% of phenotypic variation. Based on MM analysis. Window for discovery =10cM.

6.5 Conclusion

The objectives of the power simulations were to compare methods, estimate the likely success of an association study in the available panels and to extrapolate beyond the existing panel sizes to suggest what might be achieved with more markers and/or individuals.

It is evident that in these highly structured datasets GC is only reliably effective as a means of controlling the false positive rate, since this is achieved at the expense of power to detect all but the most highly associated loci. Since the most highly associated loci in naïve association do not necessarily correspond either to true discoveries nor to a full set of possible discoveries the loss of power is unnecessarily large.

PCA and SA both include mathematical models of the population structure as covariates in regression. Both achieve similar control of false discovery and retain similar power. However, two deficiencies emerged in these analyses. Firstly, PCA and SA do not fully control false discovery since they do not return the majority of marker-trait associations to the null in a P-P or Q-Q plot. Secondly, there is, at least, circumstantial evidence that this uncontrolled inflation is due to the effect of kinship since the mixed model is able to control inflation completely and the additional explanatory variable in the mixed model is kinship.

Estimates of association by SA and PCA may be subjected to GC as a way to control for residual inflation but this step is still subject to the same problems of non-specific deflation of the test statistic which cause GC itself to lose power.

The evidence presented here supports the choice of the mixed effects model for association mapping in highly structured populations. Although the mixed model is not guaranteed to give the highest power, nor yet the lowest observed false discovery rate, of any method, it does offer the most consistent balance between the two (measured as efficiency). Moreover, with one exception, the gain over the performance under the null (power ratio) is highest or joint highest wherever there is a meaningful gain. The exception being the case of GC which occasionally finds a very highly associated marker for a single locus trait and so records a low number of true discoveries, this results in low power but high power ratio.

Turning to the question of likely success of association mapping in these populations, there is little doubt that major gene effects will be usefully detected with small panels of both markers and individuals. Figure 6-24 shows that a panel of only 150 individuals and 500 markers is sufficient to give ~75% power using the mixed model in the simulated cereal.

In barley **Figure 6-19** shows ~95% power to detect major gene effects in the real panel. In the much larger wheat genome the corresponding power is ~55% (**Figure 6-18**). In the case of complex quantitative traits, here modelled as being controlled by 10 loci, Figure 6-28 shows that the proportion of loci discovered will be about 65% in a panel similar to that available for barley but only about 20% for a panel of similar size to that available in wheat. For comparison Figure 6-8 shows that the power to detect the perturbation to the yield trait in wheat is about 20% when effect size is 10%, in barley a comparable figure is 80% (Figure 6-17).

7 Association Mapping: Wheat Panel.

Using the available phenotypes (specific weight, yield, protein and Hagberg falling number. See; **5.2.1**) I attempted a genome-wide association analysis using the wheat panel.

7.1 Structure and kinship

As has been discussed (Chapter **3**) population structure/kinship can potentially lead to high levels of false discovery in association mapping because it can act as a hidden variable in regression causing an apparent association between markers and phenotypes which co-vary with structure.

Before any mathematical treatment can be applied to attenuate the association signal introduced by structure, structure must itself be modelled mathematically. In this case I used a subset of available markers as a basis for defining structure in terms of (i) an admixture model using STRUCTURE, (ii) principal coordinates of a covariance matrix based on standardised genotype and (iii) a representation of kinship as genetic correlation.

A subset of markers were chosen on the basis of both minor allele frequency and independence. High between marker pair correlation can occur when groups of markers map to the same point in the genome; in this circumstance they are clearly not independent of each other. If groups of non-independent markers are included when modelling structure it is possible for small regions of the genome to have a disproportionate influence on the way in which structure is defined. An example of this was given in (**5.2.5**) where the omission of chromosome 1B from the wheat panel largely erased the population division corresponding to 1B1R.

Figure 7-1 to Figure 7-3 shows the effect of screening the panel at different values of R^2 and MAF on population structure evident in the first two principal coordinates of a genetic correlation matrix. Table 7-1 shows the marker numbers available following this screening process.

Table 7-1. Numbers of markers remaining after screening the data at various values for MAF and between marker R^2.

		Minor allele frequency ≥		
		0.01	0.05	0.1
Maximum between marker R^2	1	1294	1235	1047
	0.9	708	667	562
	0.7	579	543	458
	0.6	534	500	424

The subset of markers ultimately chosen to define structure had MAF ≥ 0.01 and the subset were also screened such that no marker pair correlated with each other $R^2 > 0.6$. This subset had already been demonstrated to give more potential power in association analysis.

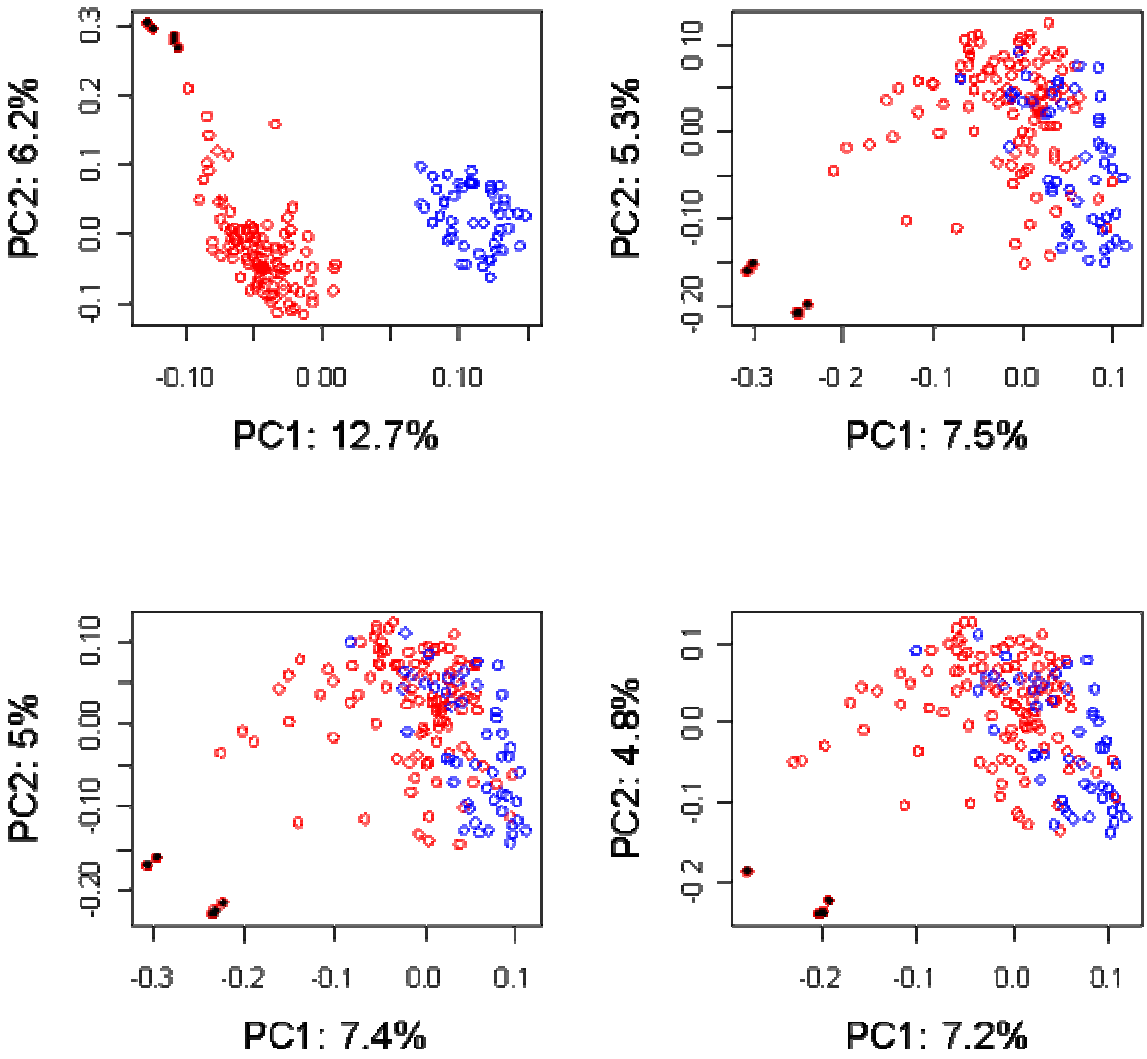


Figure 7-1. MAF is 0.01. The effect removing one marker of any pair correlated at $R^2 \geq 1.0$, 0.9, 0.7 and 0.6 (respectively, top row left to right then bottom row left to right). Plots are of the first two principal coordinates of a genetic correlation matrix. Blue circles indicate varieties with the 1B1R translocation present, red circles indicate absent. Black dots are a group of closely related facultative spring varieties.

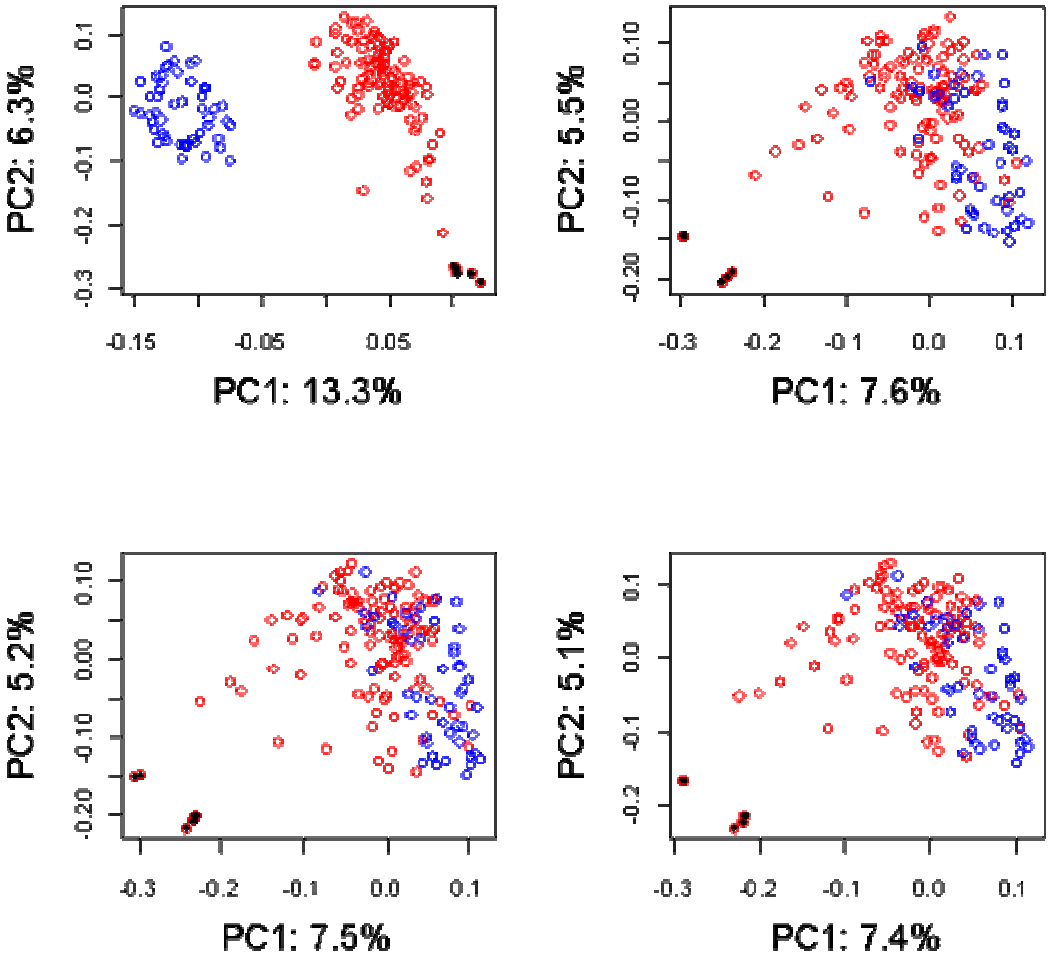


Figure 7-2. MAF is 0.05. The effect of removing one marker of any pair correlated at $R^2 \geq 1.0$, 0.9, 0.7 and 0.6 (respectively, top row left to right then bottom row left to right). Plots are of the first two principal coordinates of a genetic correlation matrix. Blue circles indicate varieties with the 1B1R translocation present, red circles indicate absent. Black dots are a group of closely related facultative spring varieties.

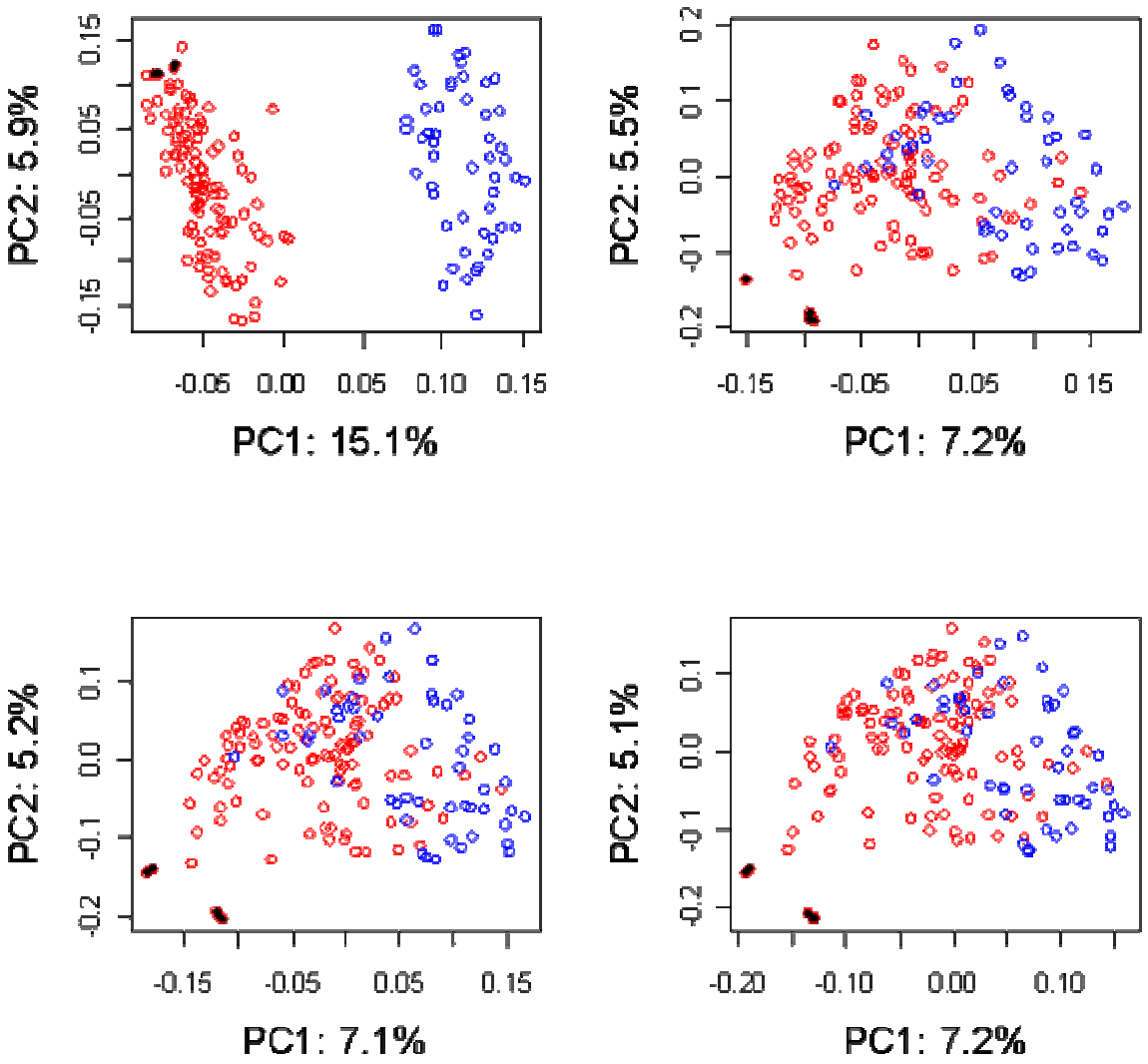


Figure 7-3. MAF is 0.1. The effect of removing one marker of any pair correlated at $R^2 \geq 1.0$, 0.9, 0.7 and 0.6 (respectively, top row left to right then bottom row left to right). Plots are of the first two principal coordinates of a genetic correlation matrix. Blue circles indicate varieties with the 1B1R translocation present, red circles indicate absent. Black dots are a group of closely related facultative spring varieties.

7.1.1 Defining the Q matrix.

The program STRUCTURE (Pritchard *et al.* 2000; Faluch *et al.* 2003, Faluch *et al.* 2007) is a valuable tool for inferring the structural origins of individuals in an association panel. The principle output is the Q matrix (**Q**) which assigns fractional membership of each of *k*

historical subpopulations to each member of the variety panel. **Q** is, therefore, a mathematical representation of substructure and the columns of **Q** may be used as co-variates in association mapping.

As has been discussed (3.2.1) the instability of estimates of **Q** and the best value of *k* are both problematic features of STRUCTURE. In this case I used a burnin of 250000 cycles and an MCMC phase of 750000 cycles for values of *k* in the range 1 – 15. All **Q** matricies were estimated in duplicate and duplicates were only accepted if the average maximum correlation between columns between duplicates was 0.999 or greater. Where this standard was not achieved for any value of *k* further STRUCTURE runs were conducted.

The suitable value of *k* was determined by considering the change in Ln(P(D)) (Figure 7-4 and Figure 7-5), the change in F_{ST} from the data over different models (Figure 7-6 and Figure 7-7) and the error accounted (adjusted R^2) in predicting phenotype using the Q matricies at each value of *k*, (Figure 7-8).

The change in Ln(P(D)) proved inconclusive since the model was still increasing in likelihood at *k*=15, an effect noted (3.3.3 and Figure 3-8). However Figure 7-6, suggests that models above *k*= 11 offer no increase in F_{ST} suggesting that adding further subpopulations does not account for further genetic variance. The examination of ΔF_{ST} and ΔK were inconclusive since, although there was a clear maximum in both cases (*k*=2 and 5 respectively) these corresponded to values of *k* where ln(P(D)) and F_{ST} were still rising strongly. There is some correspondence between Figure 7-8 and Figure 7-4 in that secondary peaks in ΔK correspond to points at which R^2 plateaus for the various traits (protein and yield around *k*=6 and HFN and specific weight around *k*=10 or even *k*=13.

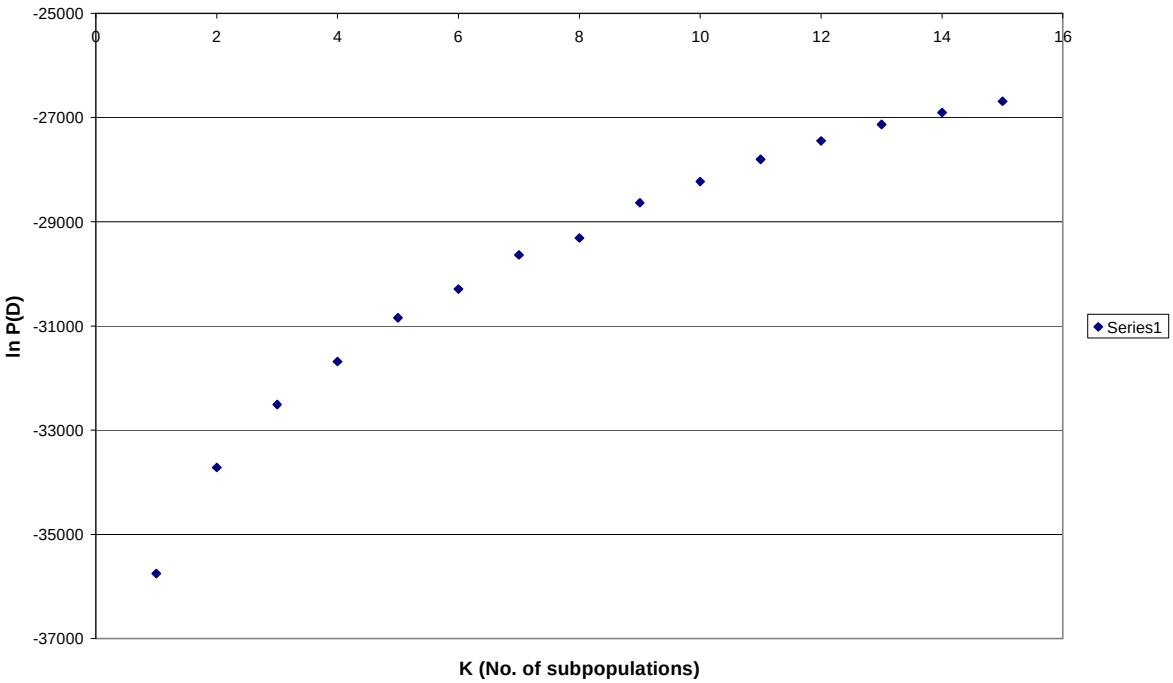


Figure 7-4. As the model of structure improves the log-likelihood of the model rises. The rising trend continues to K=15. It is not possible to infer an appropriate value of *k* from these data.

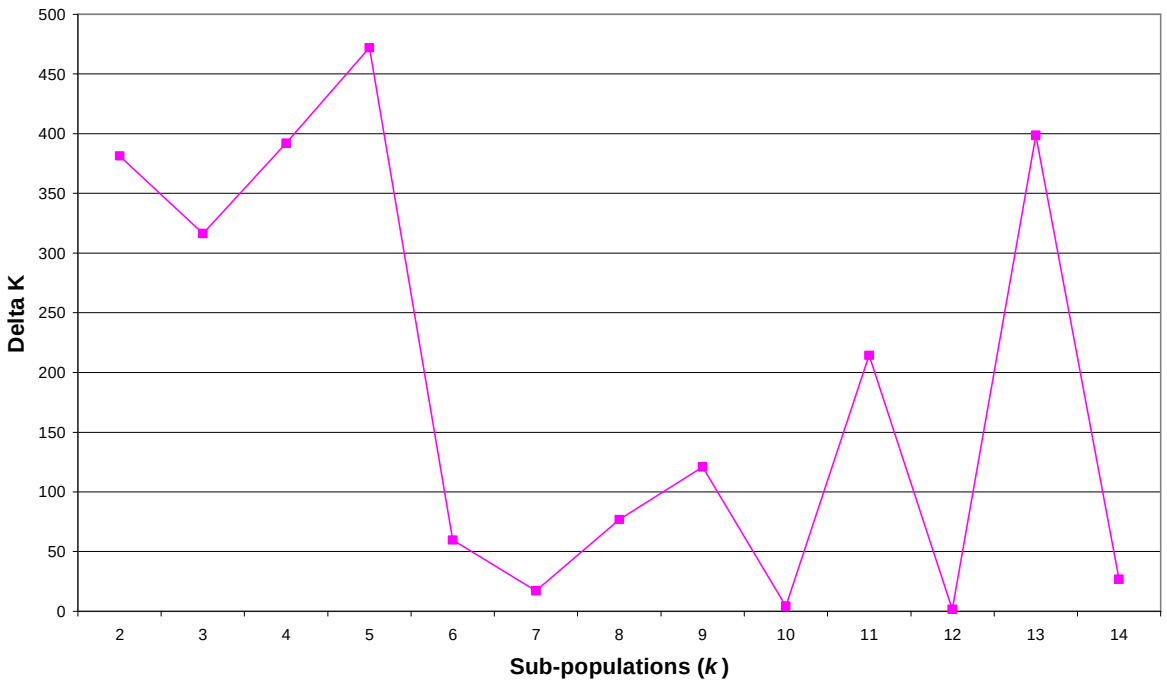


Figure 7-5. The normalised second order rate of change in Ln(P(D)) (Delta K) over number of sub-populations. The technique was expected to show a clear peak corresponding to the value of *k* which best describes the data. However the results are ambiguous.

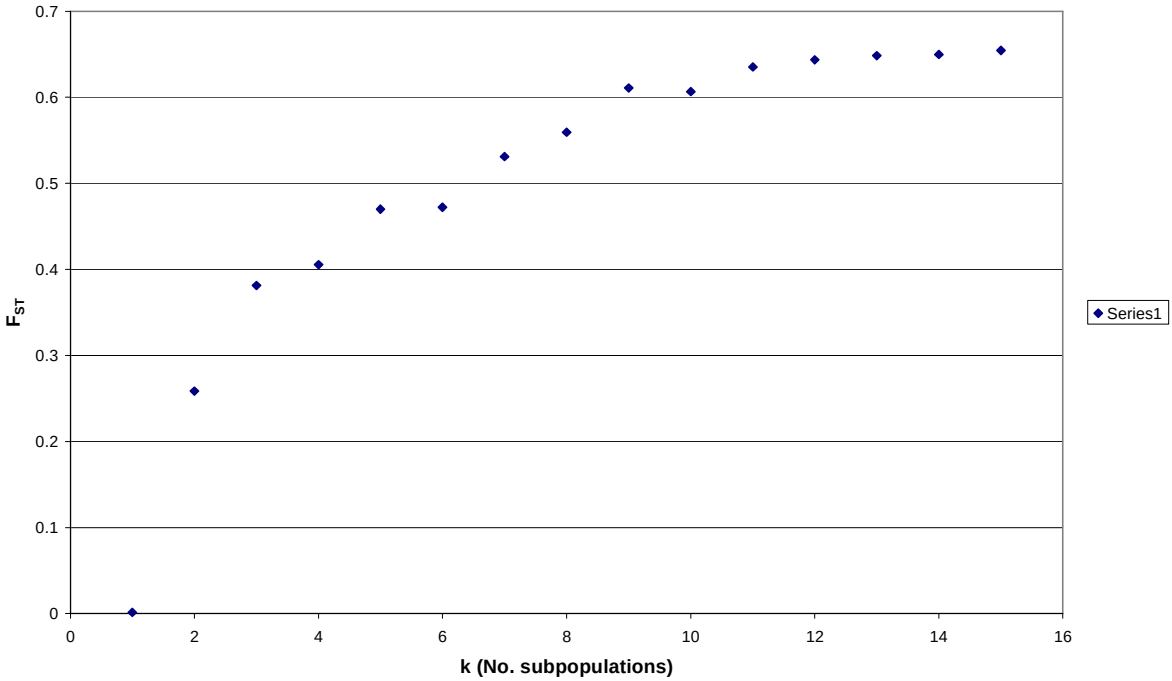


Figure 7-6. As the model of structure improves the value of FST rises. The rising trend ceases at ~ *k* =10 suggesting that above *k* =10 additional subpopulations are failing to account for additional genetic variation. This suggests **Q** based on *k* =11 would be an appropriate model for structure.

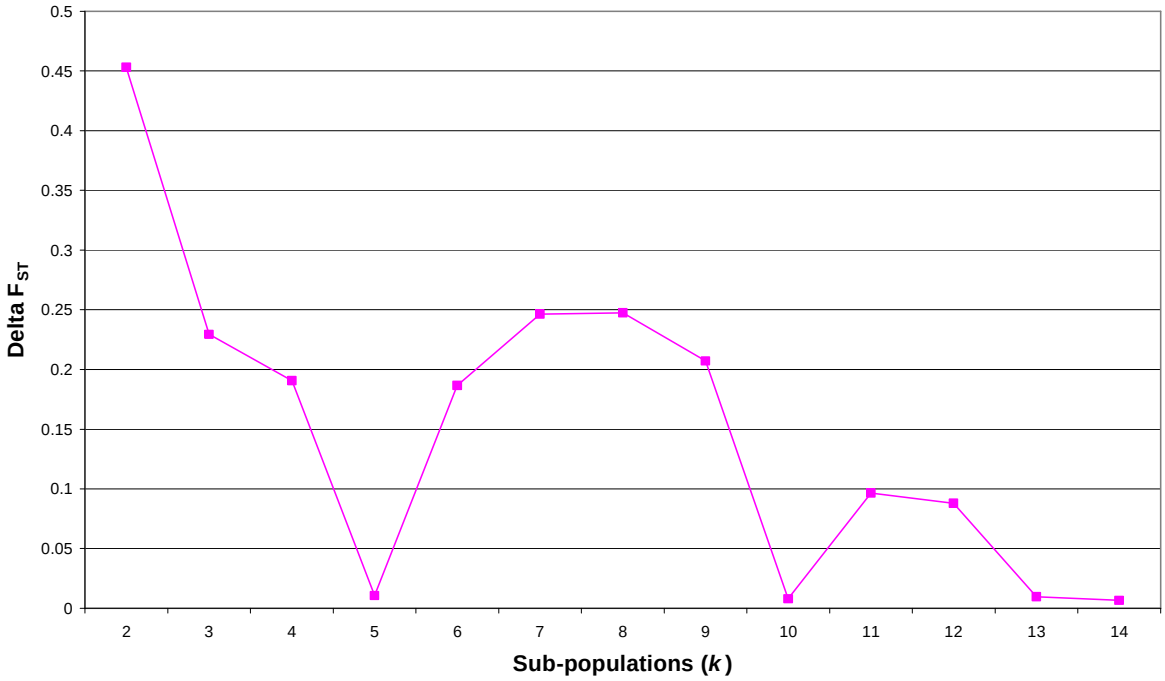


Figure 7-7. The normalised second order rate of change in F_{ST} (Delta F_{ST}) over number of sub-populations. The technique was expected to show a clear peak corresponding to the value of *k* which best describes the data. However the results are ambiguous.

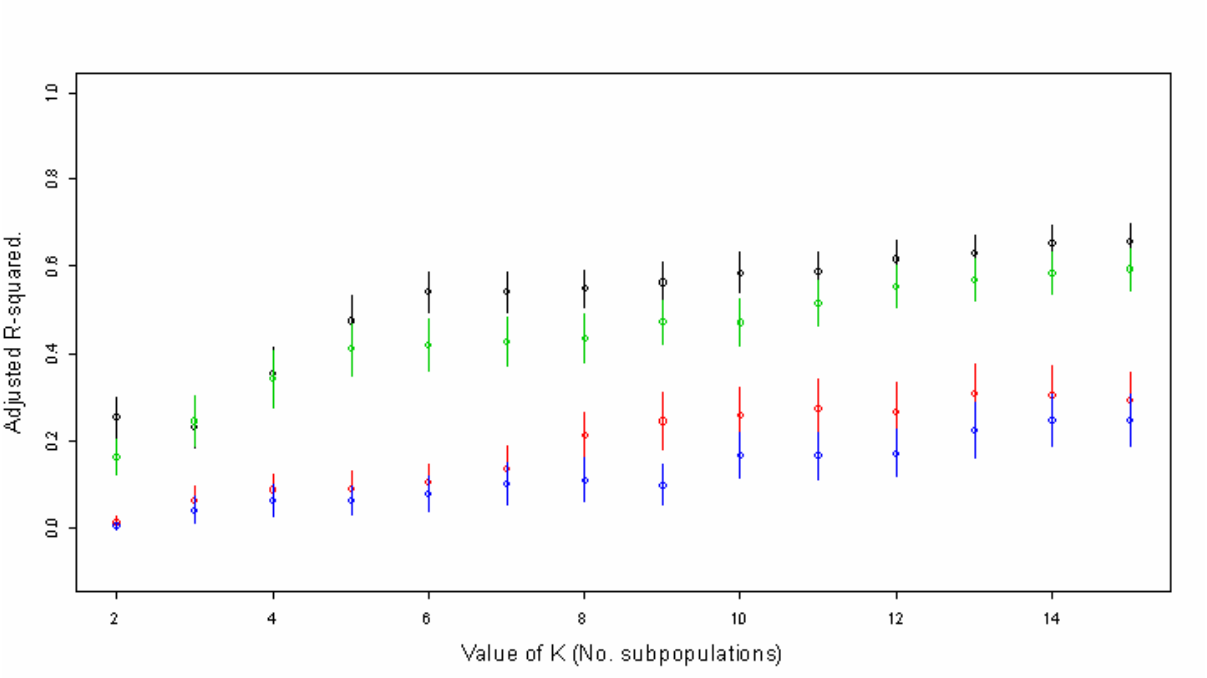


Figure 7-8. Adjusted R^2 calculated for linear models in which (*k*-1) columns in the **Q** matrices are used as explanatory variables to predict phenotype. As the model of structure improves the amount of phenotypic variation which can be explained by the **Q** matrix increases. For each phenotype there is a value of *k* above which R^2 appears not to rise, this value is not the same for all phenotypes. Black=yield, red=HFN, green=protein, blue=specific weight. (Note; the **Q** matrix provides *k* covariates but only K-1 may be used in any regression model because the fractional memberships for any individual variety sum to unity. For this reason *k*=1 is not a valid model for this analysis).

It is not unexpected to find that the minimum value for *k* varies with phenotype, this merely reflects the fact that some phenotypes are not associated with all features of the population structure, and for this reason too the correspondence between Figure 7-8 and Figure 7-4 is probably coincidental. In these data the appropriate value for *k* is probably best defined in terms of explanation of phenotype since ultimately this is the factor which will lead to spurious inflation in association analysis.

7.1.2 Defining structure in terms of principal coordinates.

As has been discussed (3.2.1.1) a commonly used method of modelling structure is in terms of a sub-set of the principal coordinates of a similarity matrix. Price *et al.* (2006) suggest the first 20 eigenvectors of a covariance matrix calculated on standardised genotypes are usually a sufficient model for use in association mapping, however there is some discretion available

to the analyst in this regard, although clearly the number of vectors must be constrained since if all were used the prediction of phenotype would be perfect.

Plotting phenotypic variation explained by linear models comprising increasing numbers of eigenvectors (Figure 7-9) shows a similar pattern to that observed using vectors of the Q matrix; protein and yield variations are rapidly accounted in the early vectors. HFN and specific weight, are less well accounted, presumably because these phenotypes do not vary in concert with structure, however there is some evidence that the 10th eigenvector contains information associated with specific weight and the 12th with HFN.

Potentially, one way to determine the appropriate number of eigenvectors is to consider the gain in variation explained as successive eigenvectors are added to the model; each additional vector will explain additional phenotypic variation, if the gain exceeds the gain in explanation of genetic variation it suggests a strong association with a structural element of the population. Conversely, if the gain is less than the gain for genetic variation explained it suggests the structural element captured in the eigenvector is not strongly associated with phenotype. One difficulty with this approach is that the higher order vectors are only explaining residual error, at this point the gain from each successive eigenvector, which can differ markedly from the last in either direction, represents a process of over-fitting and is of no value for modelling structure. Figure 7-10 shows this gain in information, a moving average has been used to aid interpretation. The figure shows that for protein and yield the early vectors provide more information about phenotypic variation than they do about genetic variation, which is consistent with Figure 7-9. Variation in specific weight and HFN are most notably explained by vectors around 11 and 14.

Overall the data suggest using eigenvectors 1-10 as a model for structure would be appropriate for all phenotypes except HFN where vectors 1-14 might provide a better model.

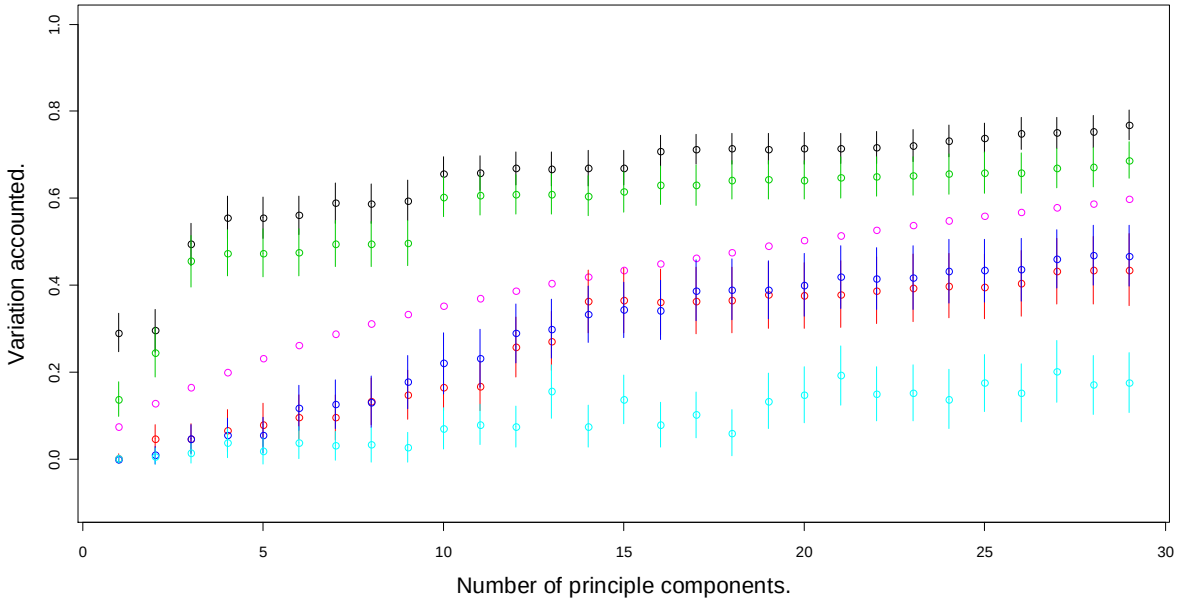


Figure 7-9. The amount of genotypic variation (mauve circles) accounted by increasing numbers of eigenvectors compared with the corresponding amount of phenotypic variation accounted. Black = yield, green = protein, blue= specific weight, red = HFN, turquoise = randomly assigned phenotype values. Error bars are ± 1 standard deviation of R^2 estimated by re-sampling the data 1000 times with replacement. As the number of eigenvectors used increase the amount of variation which can be explained by the model rises. Variation in yield and protein appears to be explained better by the early eigenvectors than does genetic variation. HFN and specific weight, however, appear to be poorly predicted by the eigenvectors and behave more similarly to the random phenotype.

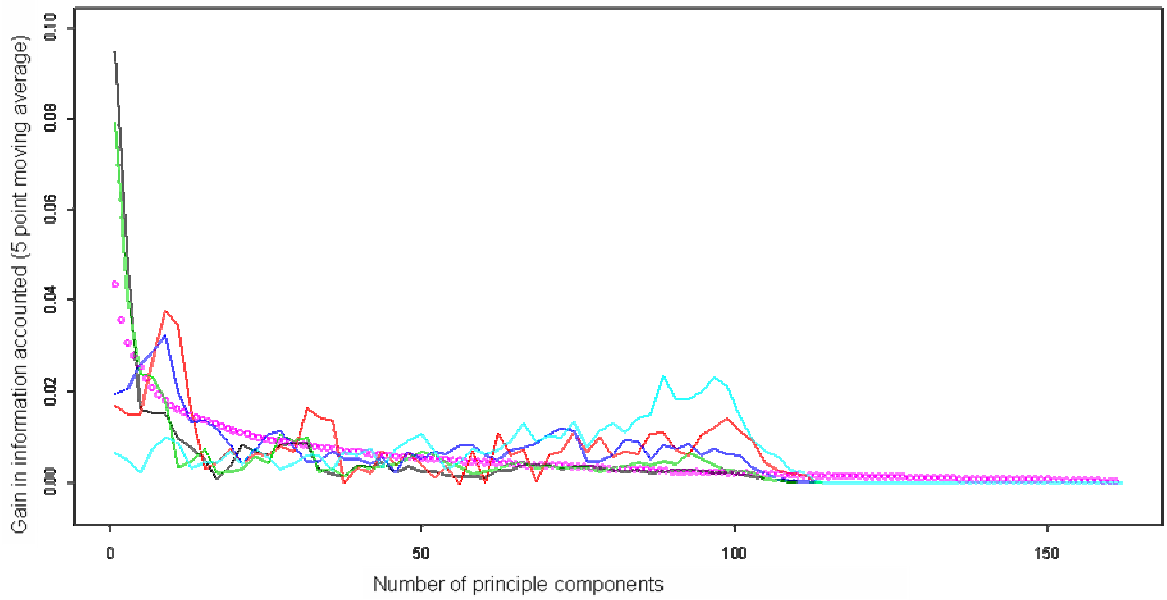


Figure 7-10. The gain in information obtained as each eigenvector is added to the model (as difference in r^2)over the number of eigenvectors shows variations in protein and yield (green and black) are being accounted for faster than genetic variation (mauve circles) in the first ~10 eigenvectors. Variation in HFN and specific weight (red and blue) are initially poorly accounted for but at ~ 10 eigenvectors there is a notable gain above the corresponding value for genetic variation. The randomly assigned phenotype(turquoise) is poorly explained by the lower order eigenvectors. Above ~120 eigenvectors there is no additional gain in information suggesting that the varieties are not independent of each other i.e. the rank of the correlation matrix is less than the length of its sides.

7.1.3 Modelling Kinship

Using genetic correlation (Astle, and Balding, 2009) calculated from markers (MAF $\geq$ 0.01 and between marker R2 $\leq$ 0.6) should better capture the gradients in inter-relationship between the varieties, including the distortions induced by back-crossing and selection, which could only have been approximated from available pedigrees (3.3.1).

A simple metric for assessing the extent to which these kinship relationships are associated with phenotype is to look for correlation between kinship and phenotypic distance (pairwise difference in phenotype) between individuals. Such correlation is a potential cause of spurious association in association mapping (Figure 7-11). In this case it is evident that yield

and protein are significantly associated with kinship, HFN is associated but considerably less strongly and variation in specific weight appears almost independent of kinship.

Evidently for protein, yield and possibly HFN it may be necessary to include kinship as part of a mixed model. However, it is interesting to note that this finding (that protein and yield are potentially susceptible to spurious association with population structure) is consistent with the findings in both 7.1.1 and 7.1.2 which is consistent with the view that the genetic correlation matrix encompasses both structure and kinship.

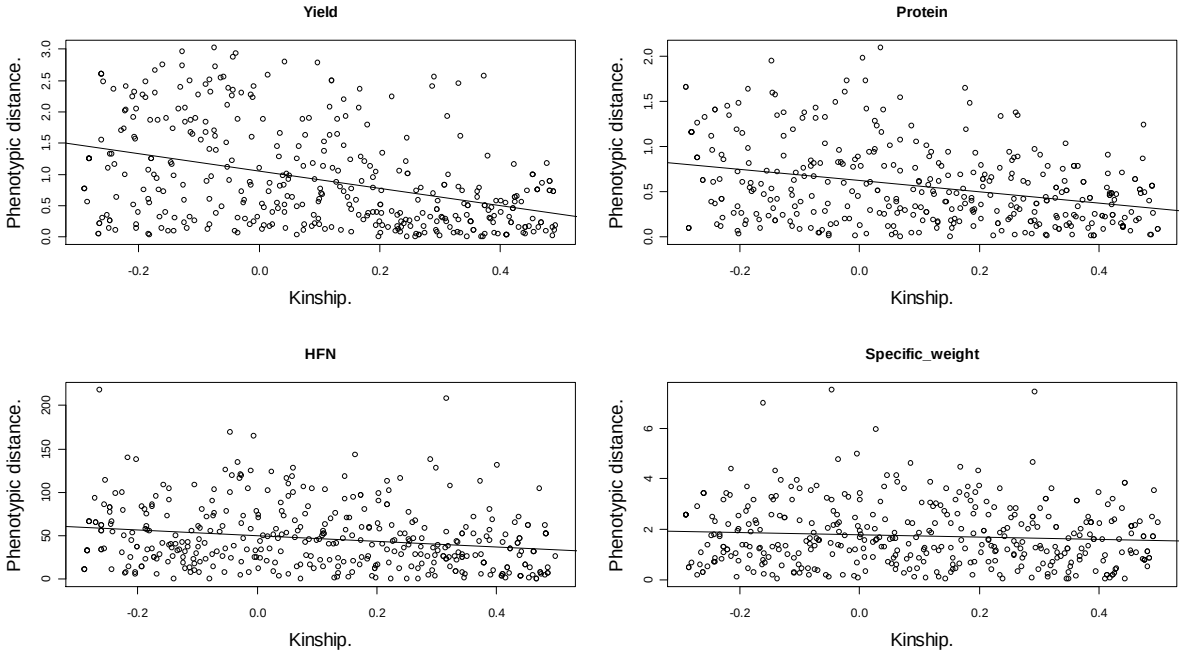


Figure 7-11. The correlation between kinship and corresponding pairwise difference in phenotype for four wheat traits. Data have been re-sampled with replacement to give an even distribution of observations over the range of kinships. Lines are best fit by least squares. 1000 re-samplings of the data were used to obtain a distribution of p-values for correlation (null hypothesis of no association). The upper 95th percentile of this distribution is the quoted P-value. Protein and yield are significantly correlated with kinship ($P < 1x10^{-4}$, and $1x10^{-7}$ respectively but variable depending on the marker panel used to define kinship). Specific weight and HFN are not significantly correlated with kinship ($P > 0.2$).

7.2 Association Mapping.

7.2.1 Genomic Control

Association mapping, in its simplest form, is an estimate of the correlation between marker genotype and phenotype. The discussion of structure in the wheat panel (7.1) identified

association between phenotype and genetic structure in all phenotypes and particularly in protein and yield.

Nonetheless, it is a useful starting point to conduct a naïve analysis and to compare the observed result with the expectation under the null hypothesis. In Figure 7-12 PP plots are presented showing that for each phenotype there is genome-wide inflation of the test statistic and that this is most notable in protein and yield. *A priori* this suggests almost all markers are significantly associated with all the phenotypes. This is not a credible model since even if it were biologically correct the experiment is highly unlikely to have sufficient power to detect a signal dissipated over so many explanatory variables. The more likely conclusion is that this effect is due to structure.

Given that we expect only a few QTL to be detected for each phenotype the expected distribution of p-values is close to that expected under the null. If the experiment has been successful there should be a small group of markers with p-values lower than expected and these would appear as a deflection above the y=x line in the P-P plots (-log10 scale). Genomic control (GC) (3.3.5.1) can be used to return the distribution of the inflated test statistic much closer to that expected under the null and this may reveal a small residue of significantly associated markers. Here I used the observed p-values to obtain the corresponding values of chi-squared 1df by squaring the appropriate quantiles of ~N (0,1). GC was performed on these values and corrected p-values recovered.

Figure 3-9 compares the distribution of p-values of the naïve analysis with and without GC. There is little evidence in this figure of any useful association following GC however the distribution of naïve associations for HFN and specific weight do show groups of markers at low p-value which have had their significance inflated more strongly than the majority of markers; it is conceivable that among these there are some genuine associations.

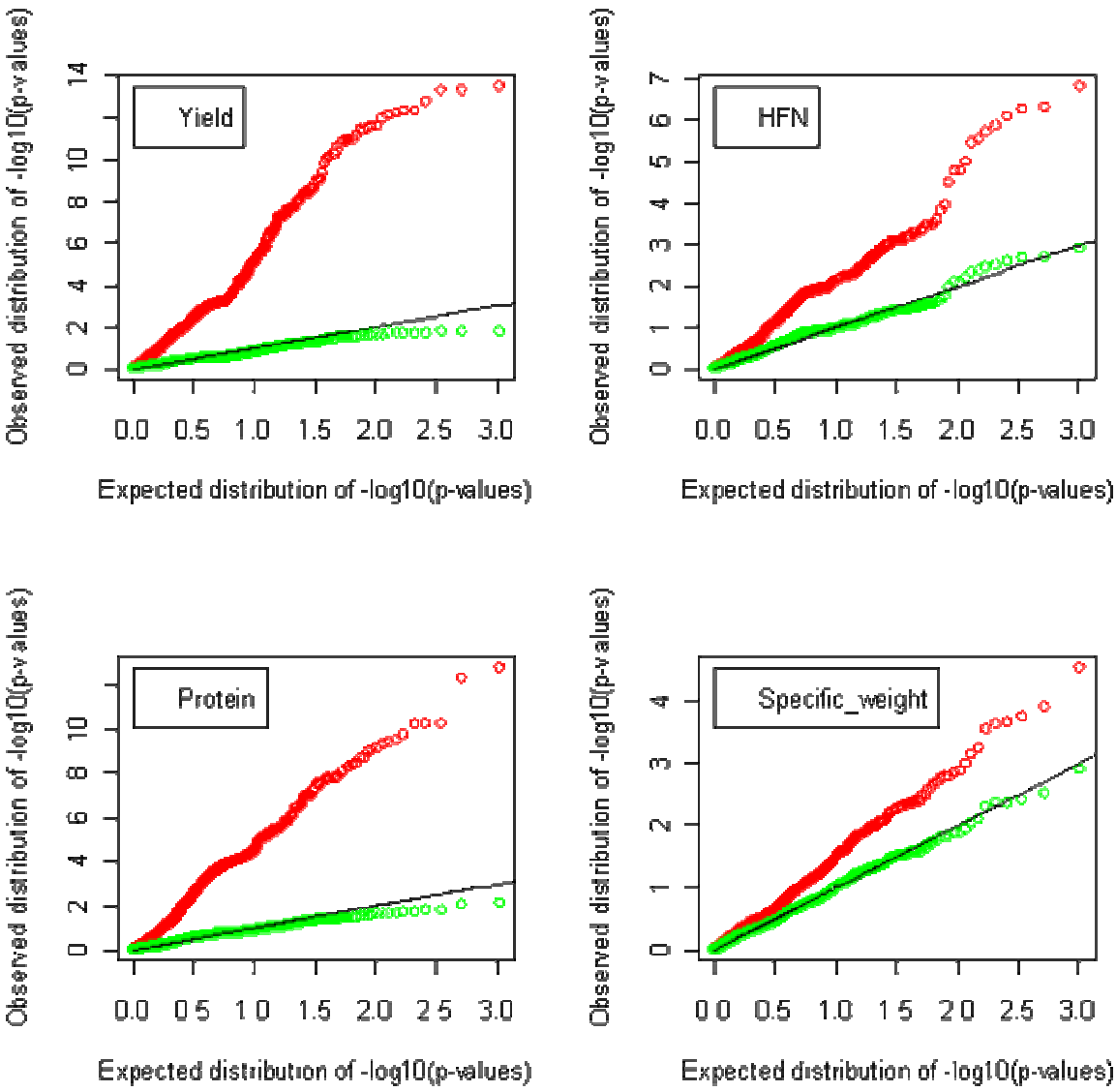


Figure 7-12. PP plots of naïve association (red) measured as correlation between marker genotype and phenotype and the same data after adjustment for inflation using genomic control (green). Note that the inflation evident for specific weight and HFN is less than that for yield and protein; the latter two phenotypes having already been shown to be strongly associated with structure.

7.2.1.1 Significance of association

Having obtained p-values for association a significance threshold must be determined, for reasons discussed (3.5.1) this can most conveniently be done by calculating the FDR (q-values) for the observed distribution of p-values. Figure 7-13 and Figure 7-14 compare the q-values obtained from p-values from naïve analysis with the corresponding results after

adjustment with GC. It is evident from this figure that adjustment with GC in these highly structured data has removed almost all signals of association with the possible exception of a small number of markers which retain an association with HFN, these are identified, mapped and summarised in

Table 7-2.

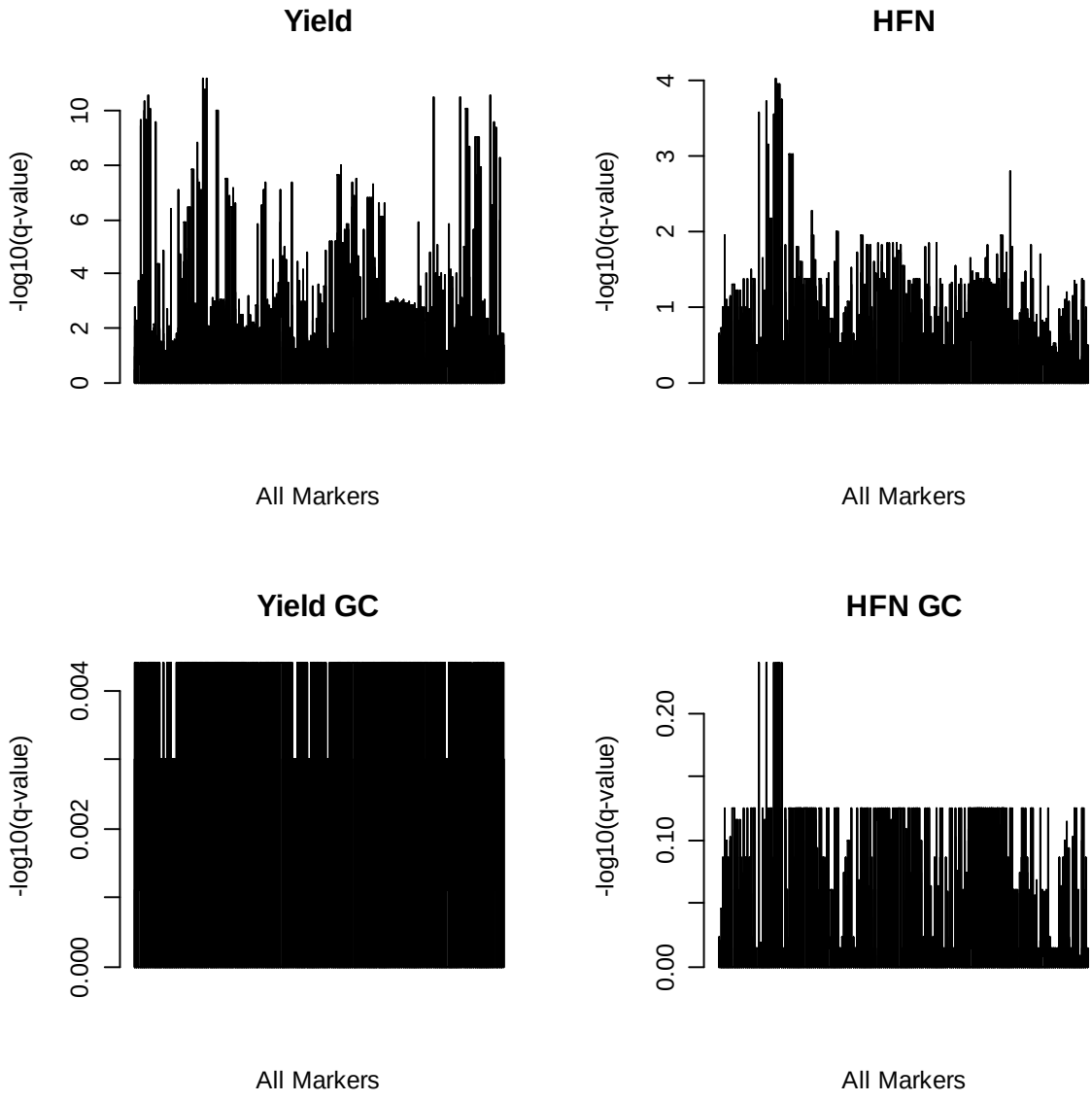


Figure 7-13. FDR (q-value) for all available markers calculated from the p-values of naïve association (upper pair of graphs) and association adjusted with genomic control (lower pair of graphs). The association signal for yield (strongly associated with structure) has been removed by genomic control whereas a very weak signal of association (FDR = 56%) persists in HFN after genomic control (association with structure is less pronounced).

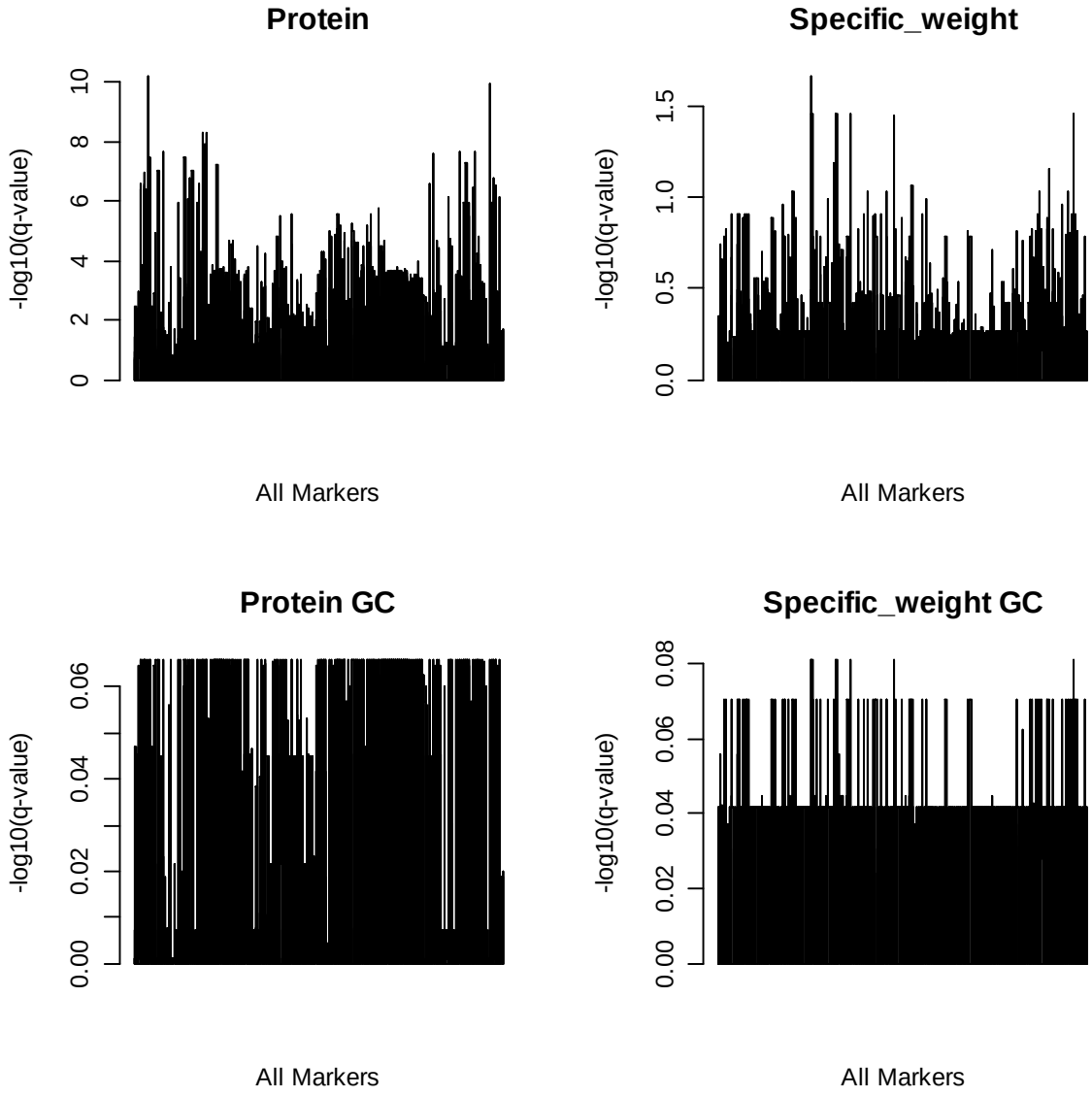


Figure 7-14. FDR (q-value) for all available markers calculated from the p-values of naïve association (upper pair of graphs) and association adjusted with genomic control (lower pair of graphs). The association signal for protein (strongly associated with structure) has been removed by genomic control, although a weak signal seems to persist for specific weight after genomic control the q-values are so low (FDR = 83%) that this should be disregarded.

It is tempting to consider the most highly associated markers following naïve association as candidates but to do so is to ignore the evidence in the P-P plots, that almost all the markers appear associated with trait and the experiment is too small to have genuinely detected this. For illustrative purposes I summarise the ten most highly associated markers for each trait on

the basis of naïve analysis (Table 7-3 to Table 7-6) these may be compared with markers identified subsequently.

Table 7-2 Markers showing association with Hagberg falling number following adjustment using genomic control.

Marker Name	Chromosome assignment[1]	cM	-log10(q-value) Genomic Control
wPt-9813	7B	84.6	0.240078
wPt-6320	7B	84.6	0.240078
wPt-1533	7B	85.2	0.240078
wPt-5069	7B	85.6	0.240078
wPt-4644	7B	unmapped	0.240078
wPt-0126	7B	unmapped	0.240078
wPt-3785	7B	unmapped	0.240078
wPt-4319	7B|7D	unmapped	0.240078

Table 7-3 Ten markers most associated with yield in a naïve association analysis.

Marker name	Chromosome assignment[1]	cM	p-value
wPt-3754	2B	unmapped	3.07E-14
wPt-0948	2B	unmapped	4.44E-14
wPt-2989	2B	unmapped	5.04E-14
wPt-6108	2B	unmapped	1.69E-13
tPt-9065	2B	unmapped	6.45E-13
wPt-7340	3A|3B	unmapped	4.49E-13
wPt-8355	3B	unmapped	4.57E-13
wPt-4386	6B|7A	unmapped	1.1E-12
wPt-2964	6B|7B	unmapped	5.62E-13
wPt-0991	7A	unmapped	2.36E-12

[1] Provided by Triticarte, the genotyping sub-contractor.

Table 7-4 Ten markers most associated with HFN in a naïve association analysis.

Marker name	Chromosome assignment[1]	cM	p-value
wPt-0126	7B	unmapped	1.45E-07
wPt-3785	7B	unmapped	4.48E-07
wPt-6320	7B	84.6	5.23E-07
wPt-9813	7B	84.6	7.91E-07
wPt-1533	7B	85.2	1.79E-06
wPt-5069	7B	85.6	2.9E-06
wPt-4644	7B	unmapped	3.58E-06
wPt-9013	7B	unmapped	9.72E-06
wPt-9515	7B	unmapped	1.63E-05
wPt-4319	7B|7D	unmapped	1.38E-06

Table 7-5 Ten markers most associated with specific weight in a naïve association analysis.

Marker name	Chromosome assignment[1]	cM	p-value
wPt-9792	1B	unmapped	2.82E-05
wPt-2461	1B	unmapped	0.000211
wPt-2315	1B	19.2	0.001319
wPt-0773	3B	unmapped	0.000281
wPt-11407	3B	unmapped	0.000721
wPt-9824	4B|7A	2.7	0.001023
wPt-1241	6B|7A	35.8	0.000178
wPt-2293	unassigned	unmapped	0.000125
wPt-7049	unassigned	unmapped	0.000229
wPt-10311	unassigned	unmapped	0.000588

[1] Provided by Triticarte, the genotyping sub-contractor.

Table 7-6 Ten markers most associated with protein in a naïve association analysis.

Marker name	Chromosome assignment[1]	cM	p-value
wPt-3754	2B	unmapped	5.33E-11
wPt-2989	2B	unmapped	5.51E-11
wPt-0948	2B	unmapped	5.96E-11
wPt-6108	2B	unmapped	1.91E-10
tPt-9065	2B	unmapped	6.66E-10
wPt-7340	3A|3B	unmapped	1.53E-13
wPt-8355	3B	unmapped	5.31E-13
wPt-6585	6B|7A	unmapped	4.33E-10
wPt-2964	6B|7B	unmapped	4.75E-10
wPt-6191	unassigned	unmapped	3.54E-10

7.3 Structured Association (SA)

The initial examination of R^2 between the Q matrix and the phenotypes (Figure 7-8) suggested that k=6 would be an appropriate model of structure in the case of yield and protein. Structure had little predictive power in the case of specific weight and HFN although Figure 7-8 did suggest Q with k=11 and greater predicted most of the variation, additionally Figure 7-6 suggested there was little increase in F_{ST} for structure models with k>11. Thus Q with k = 6 and Q with k = 11 were used to model structure.

Figure 7-15 to Figure 7-18 confirm the observation that HFN and specific weight are not explained by structure; the P-P plots for these phenotypes do not 'deflate' when the regression model is applied. The significances of protein and yield are partly deflated. In all cases genomic control, after SA, returns a distribution close to the null.

It is interesting that a residue of potentially significant markers for yield are most evident when k=6 and maximum R^2 = 1, only one potential association persists if k=11 is used. This either suggests over-correction as k increases or that k=6 is not an adequate model for structure and allows some spurious associations to persist. Given that k=6 appears to maximally explain variation in yield the former explanation is most likely.

As is the case with yield, HFN shows a group of potentially associated markers when k=6 and maximum R^2 = 1. Unlike yield, however, the effect seems to strengthen when k = 11 and the number of potentially significant markers increases, however applying genomic control to the result removes these potential discoveries. There appear to be no cases where a convincing excess of potentially associated makers is observed for specific weight or protein.

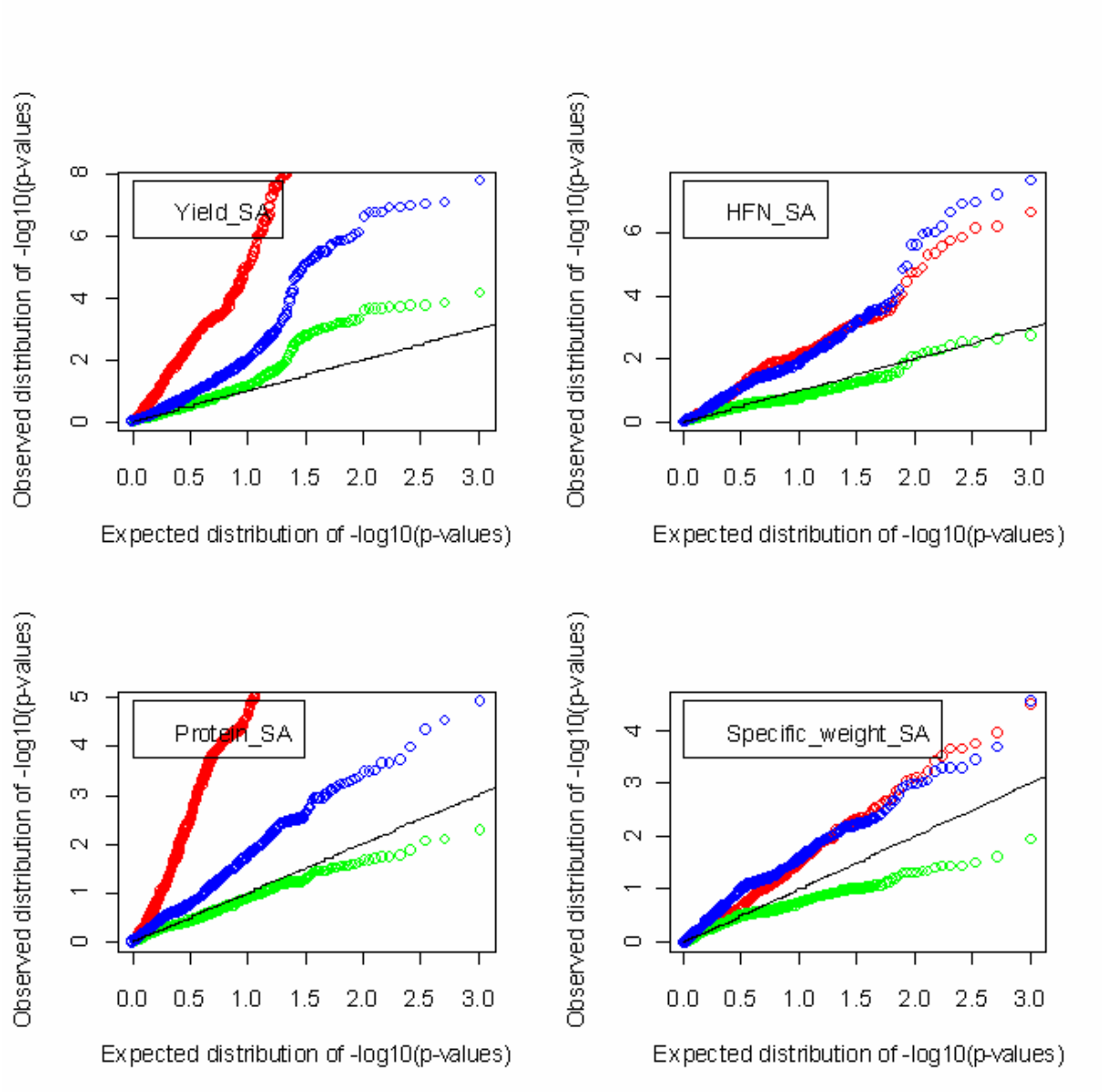


Figure 7-15. P-P Plots for the four phenotypes. The markers are restricted to MAF = 0.1 and maximum between maker R^2 is 1. Red: naïve association, Blue: association measured using structured association (Q matrix; *k*=6). Green: structured association + genomic control. Note that the inflation of the test statistic for specific weight and HFN is not affected by structured association. For this panel, there is evidence for associated markers in the case of HFN and yield.

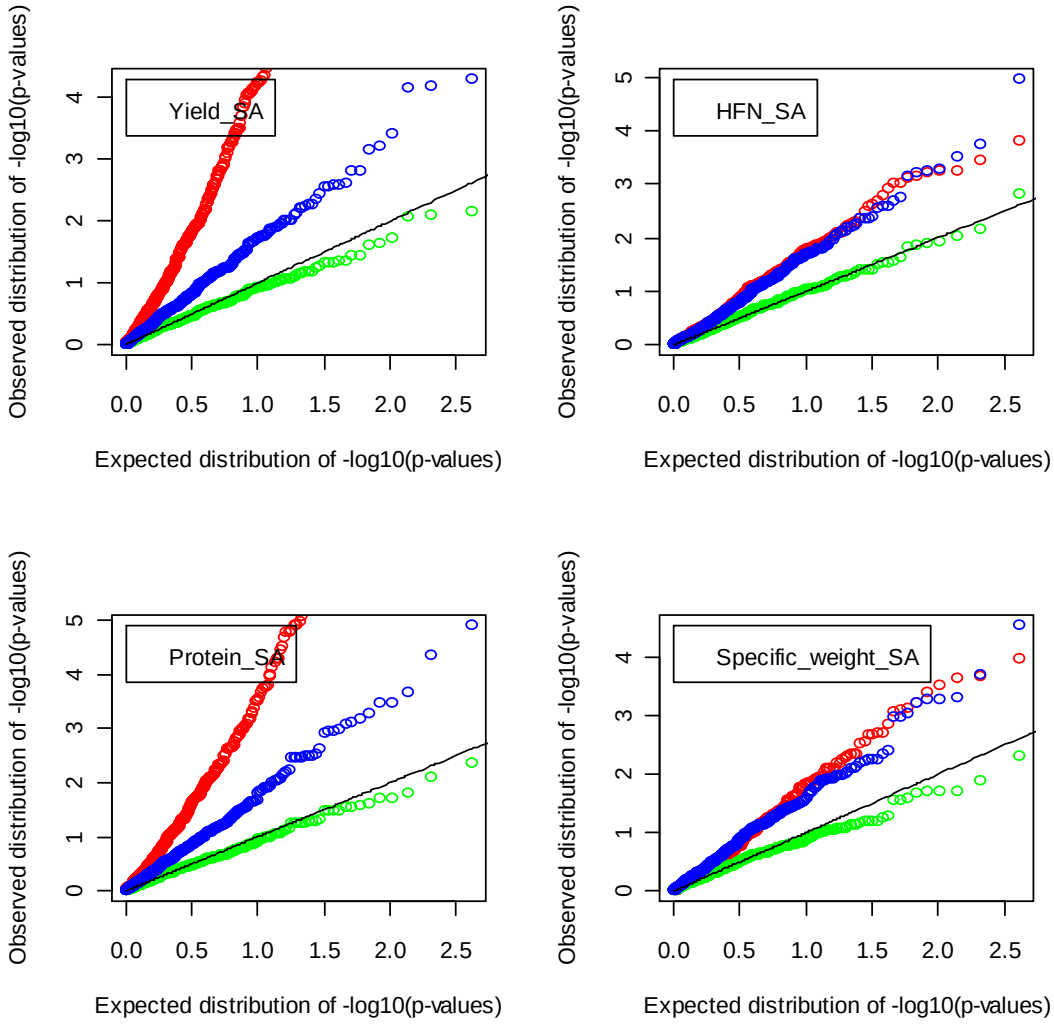


Figure 7-16. P-P Plots for the four phenotypes. The markers are restricted to MAF = 0.1 and maximum between maker R^2 is 0.6. Red: naïve association, Blue: association measured using structured association (Q matrix; *k*=6). Green: structured association + genomic control. Note that the inflation of the test statistic for specific weight and HFN is not greatly affected by structured association. There is little evidence for associated markers in this marker panel.

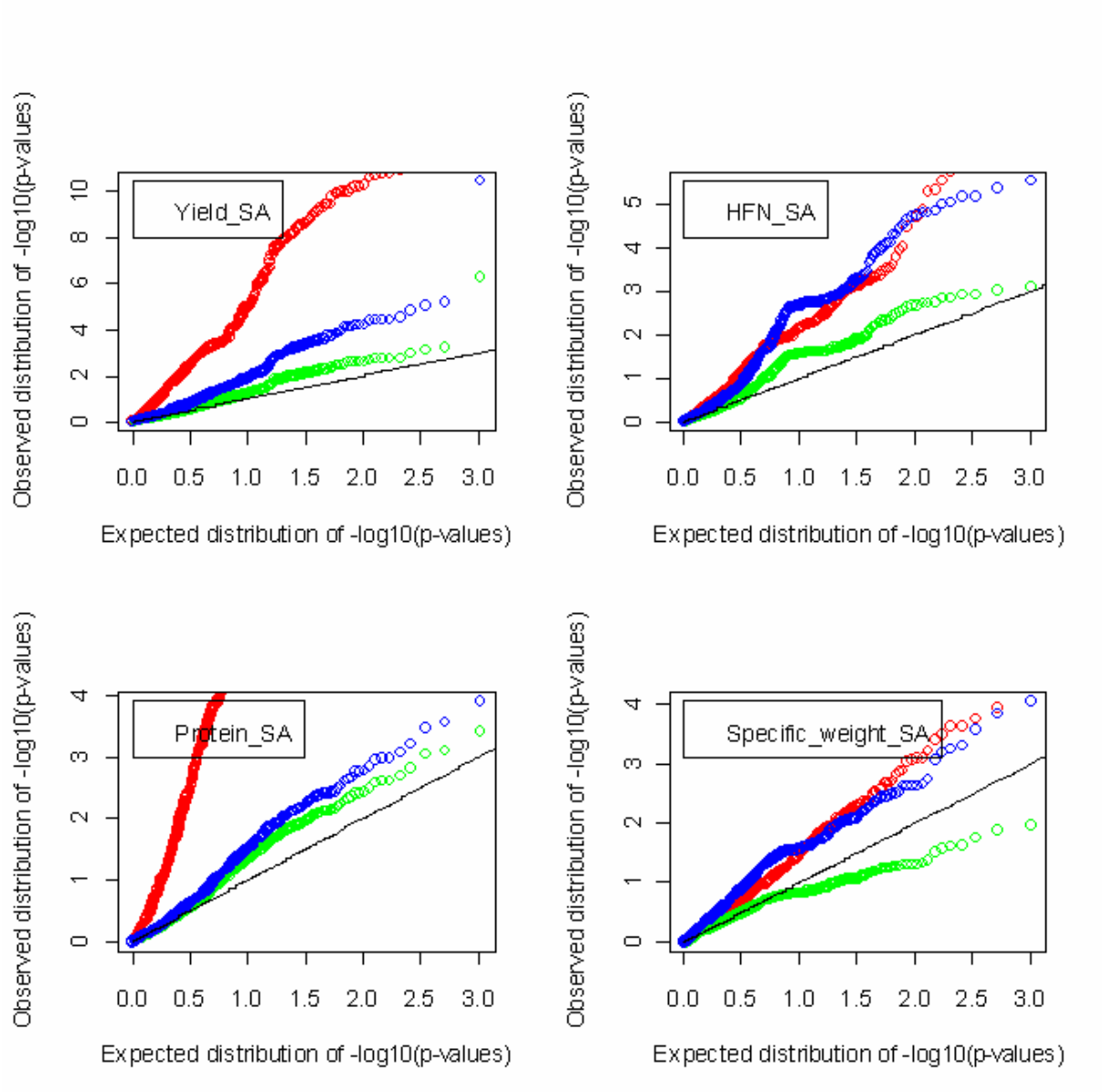


Figure 7-17. P-P Plots for the four phenotypes. The markers are restricted to MAF = 0.1 and maximum between maker R^2 is 1. Red: naïve association, Blue: association measured using structured association (Q matrix; *k*=11). Green: structured association + genomic control. Note that the inflation of the test statistic for specific weight and HFN is not affected by structured association. Some markers show association with HFN and a single marker appears associated with yield.

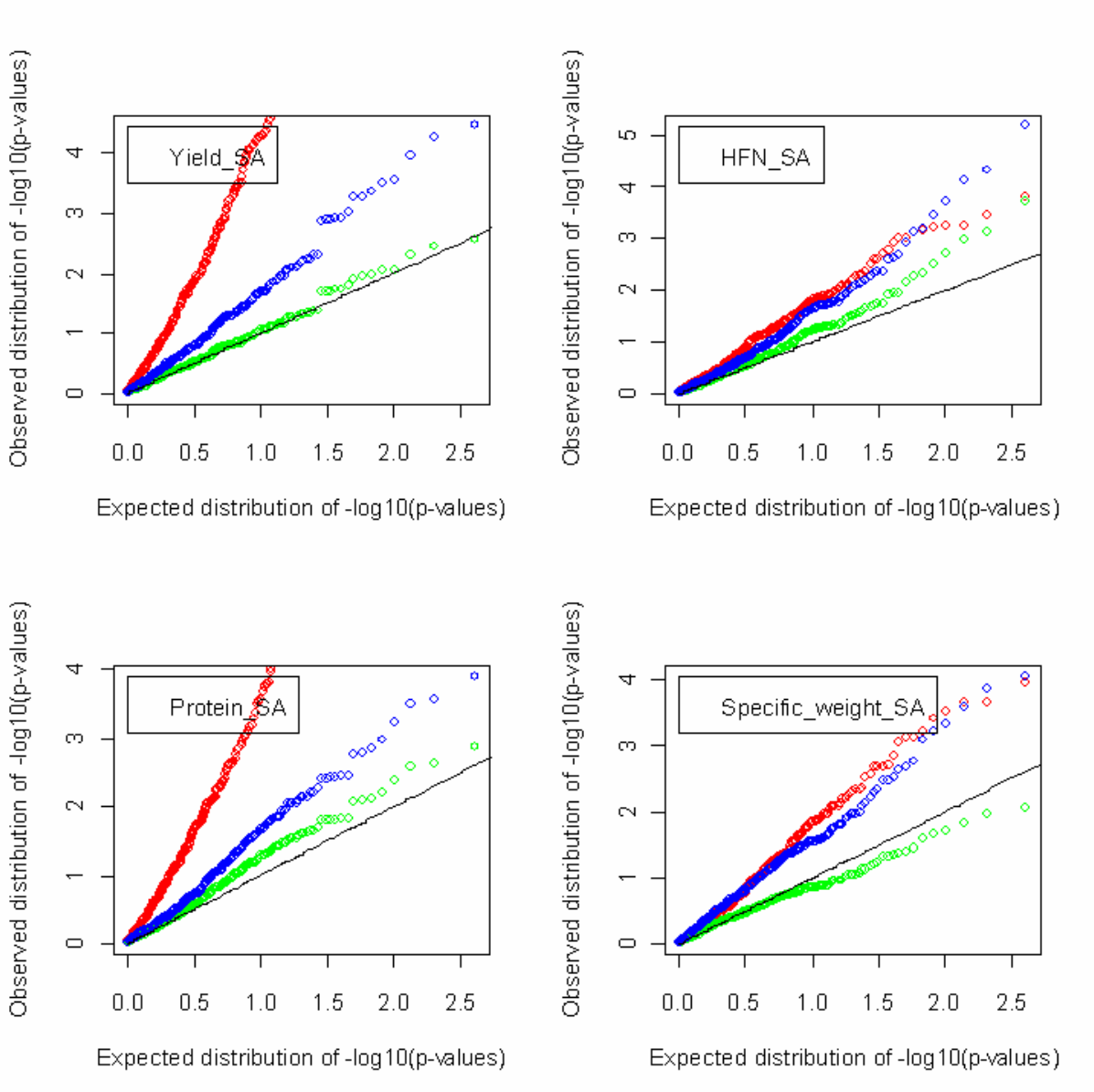


Figure 7-18. P-P Plots for the four phenotypes. The markers are restricted to MAF = 0.1 and maximum between maker R^2 is 0.6. Red: naïve association, Blue: association measured using structured association (Q matrix; *k*=11). Green: structured association + genomic control. Note that the inflation of the test statistic for specific weight and HFN is not affected by structured association. Some markers show association with HFN.

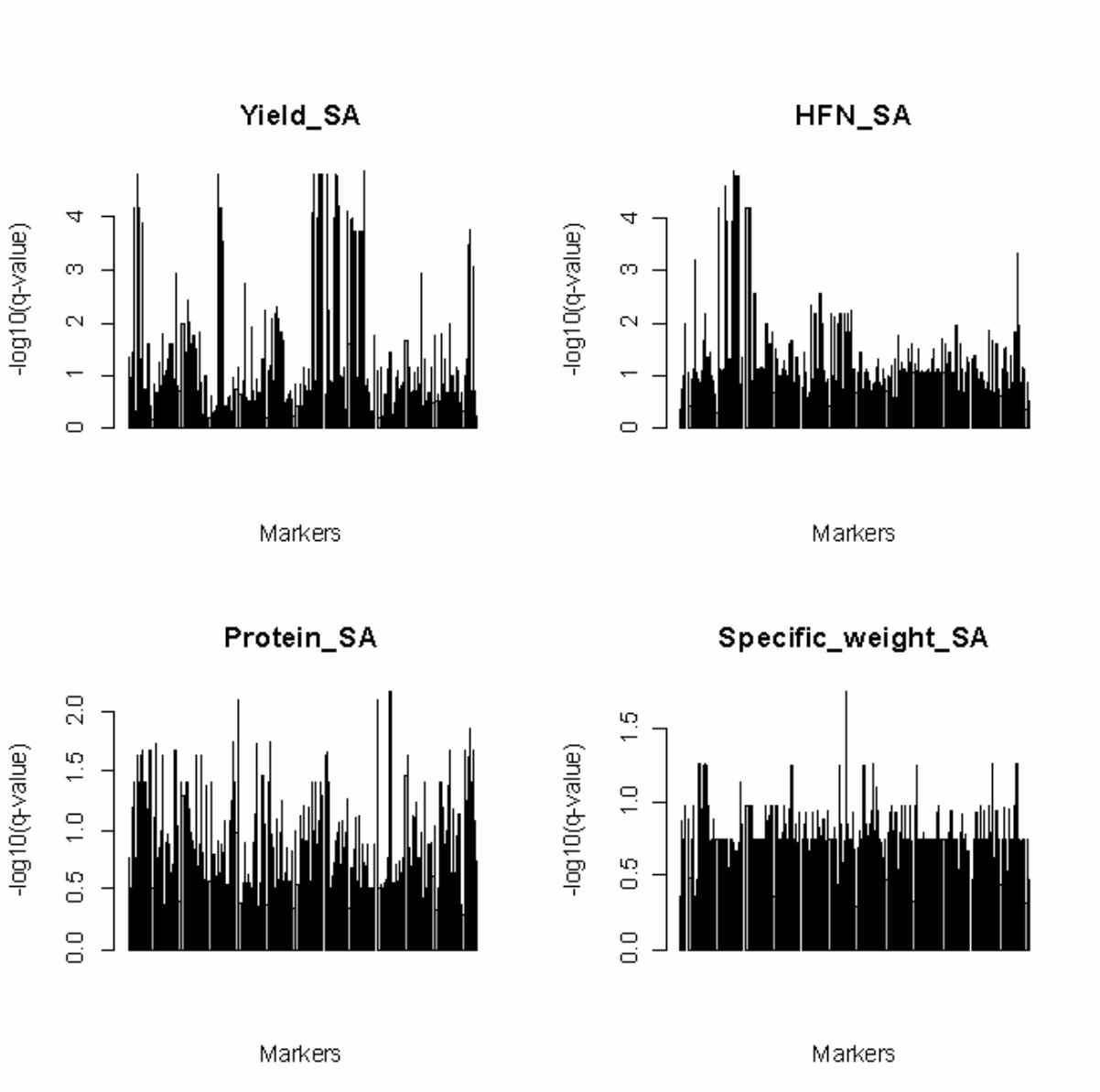


Figure 7-19. Association found following correction for structure using a Q matrix (k=6) in structured association. MAF=0.1 maximum between marker $R^2 = 1$.

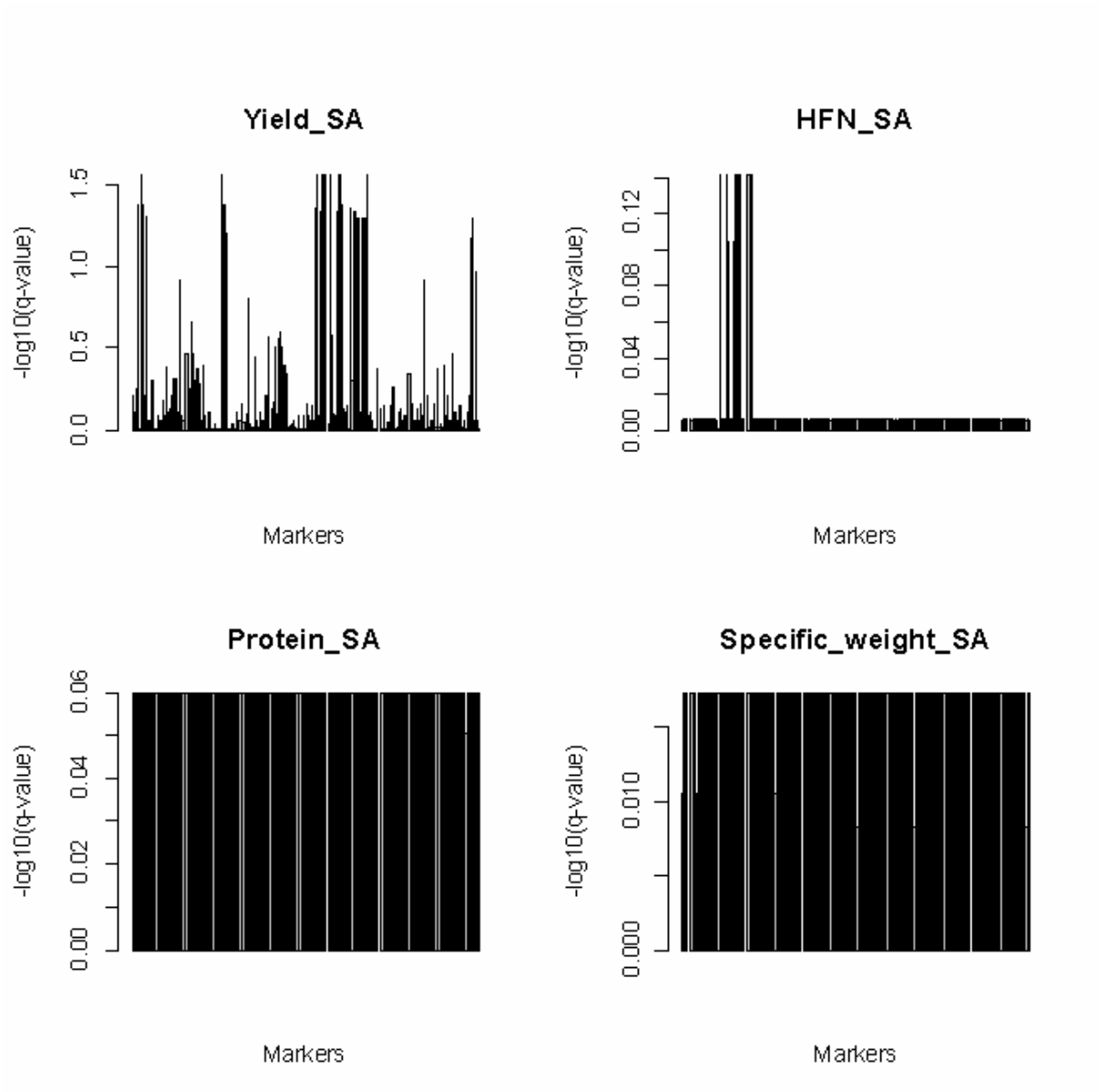


Figure 7-20. Association found following correction for structure using a Q matrix (k=6) in structured association with additional treatment with genomic control. MAF=0.1 maximum between marker $R^2 = 1$.

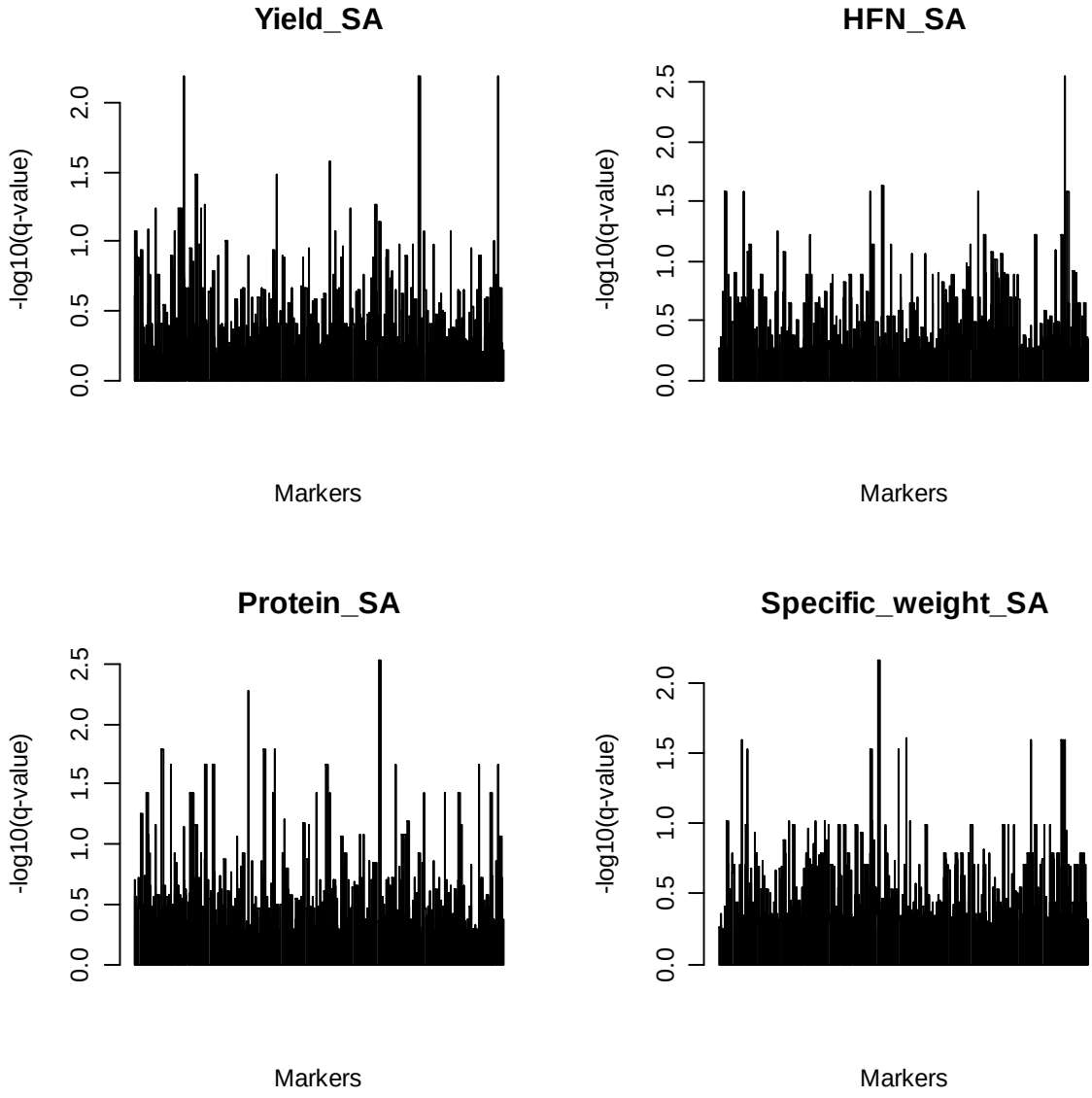


Figure 7-21. Association found following correction for structure using a Q matrix (k=6) in structured association. MAF=0.1 maximum between marker $R^2 = 0.6$.

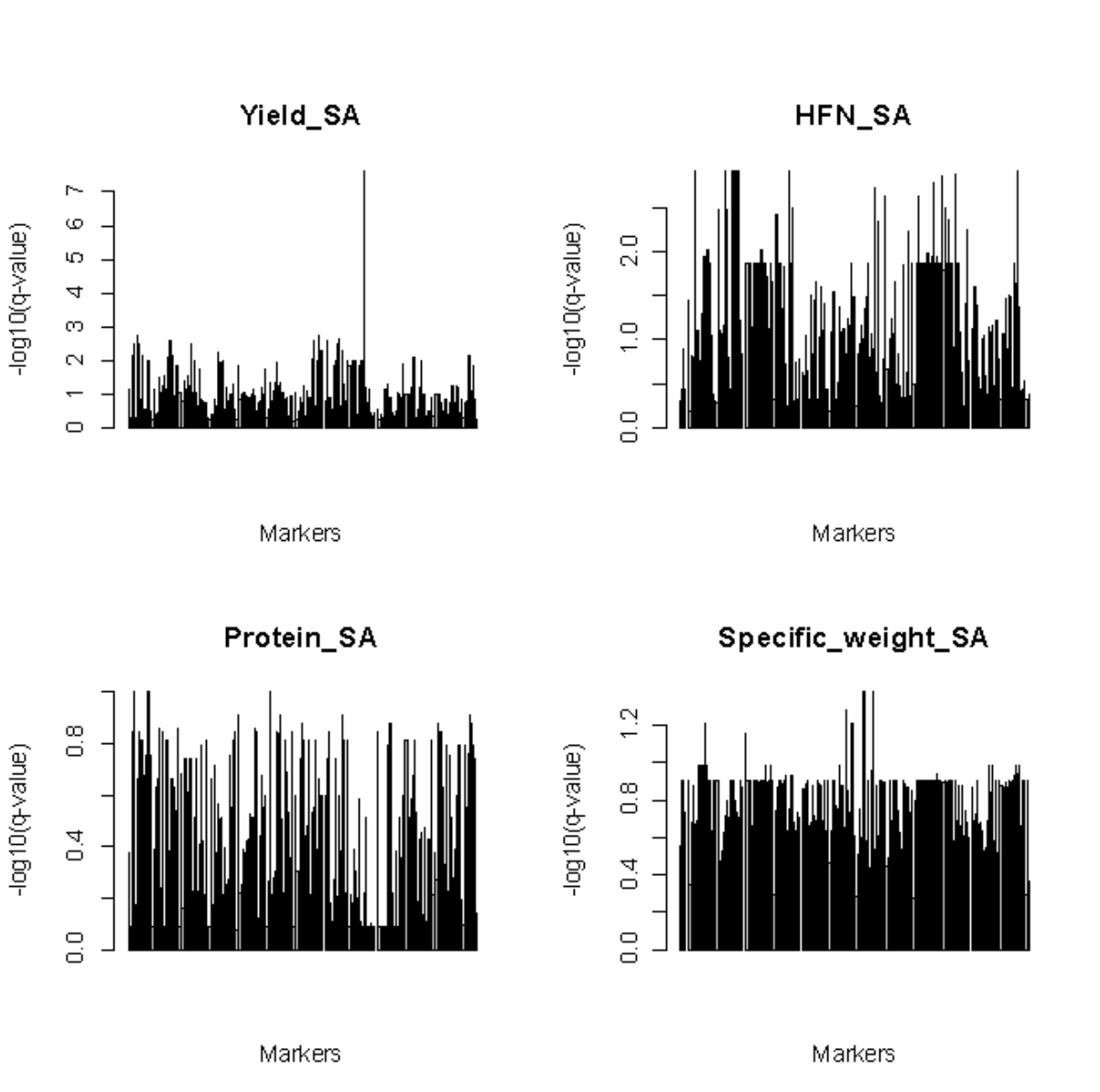


Figure 7-22. Association found following correction for structure using a Q matrix (k=11) in structured association. MAF=0.1 maximum between marker $R^2 = 1$.

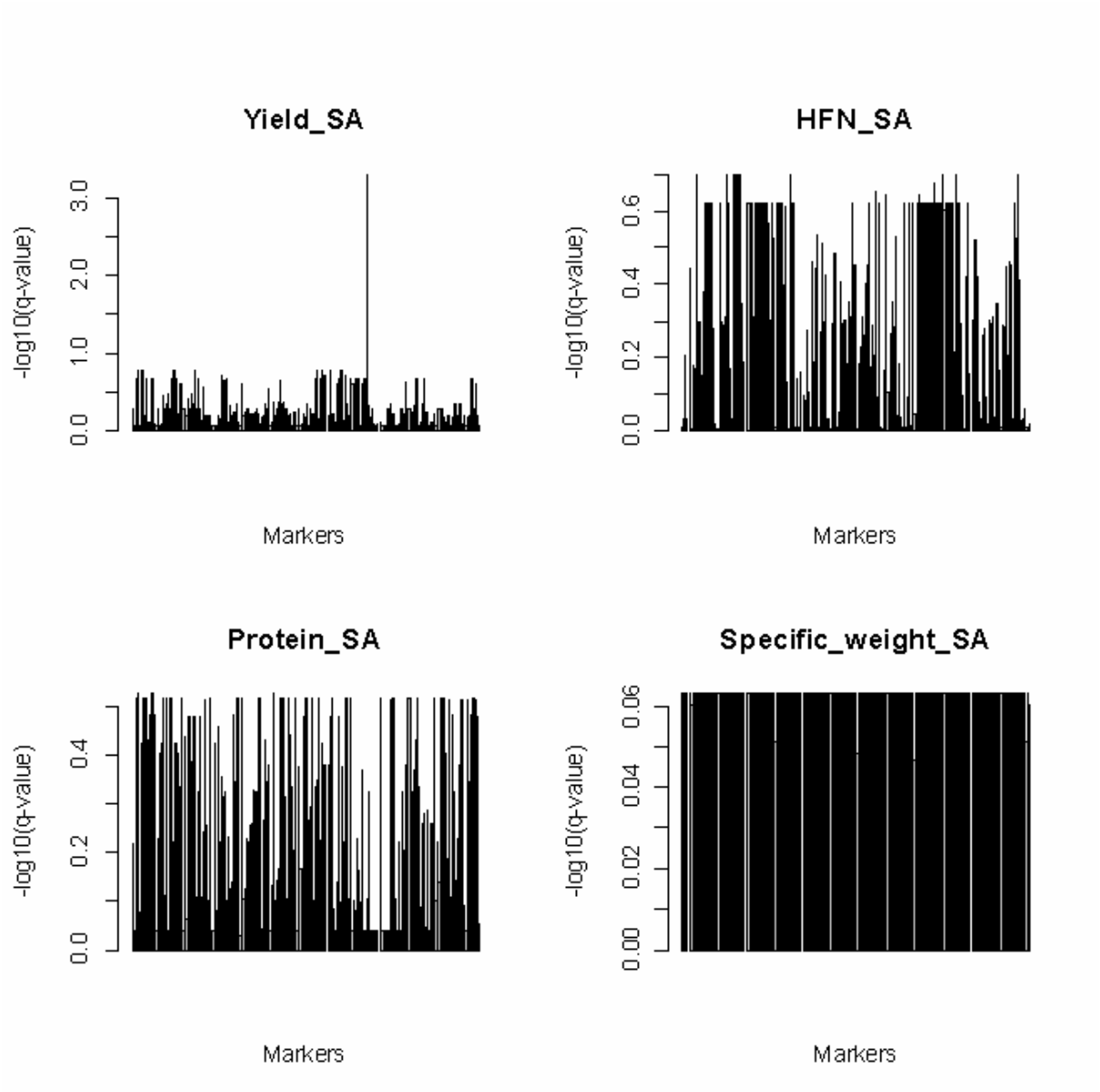


Figure 7-23. Association found following correction for structure using a Q matrix (k=11) in structured association with additional treatment with genomic control. MAF=0.1 maximum between marker $R^2 = 1$.

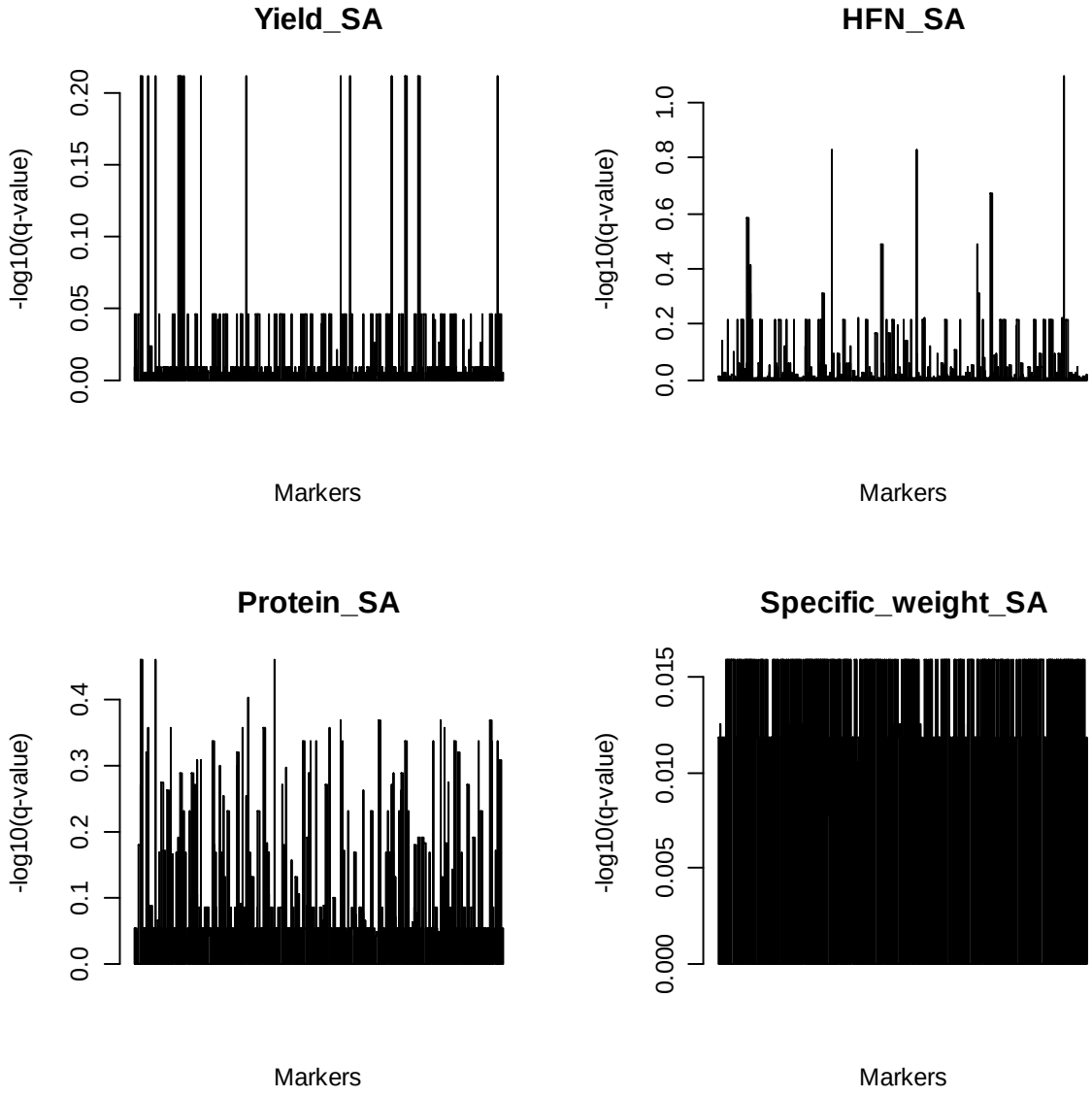


Figure 7-24. Association found following correction for structure using a Q matrix (*k*=11) in structured association with additional treatment with genomic control. MAF=0.1 maximum between marker $R^2 = 0.6$.

7.3.1 Significance of association.

Using FDR (q-values) to manage the multiple testing issue, Figure 7-19 to Figure 7-24 illustrate profiles of association for the four traits under differing models for Q, with and without GC and with the marker panel skimmed at different values of between marker R^2. The relationship between **yield** and structure was most simply defined at *k*=6 and this model results in 40 potential discoveries (FDR <0.1, *k*=6) which, due to the clumped nature of DArT

markers, corresponds to about 4 potential QTL (Table 7-8). When Q with *k*=11 is used a single marker (wPt-3467) appears far more significant than any other although the overall profile of significance of markers does not change greatly. The relationship between **HFN** and structure was best described by Q with *k*=11 and indeed when *k*=6 no discoveries are made (Table 7-7). The significance of markers associated with HFN is modest compared to those for yield but 27 markers can be identified with FDR < 0.25 (*k*=11) and an FDR of 20 - 30% may still be a useful guide to further research. Considering map position Table 7-7 suggests potentially ~ 4 QTL have been identified.

Table 7-7. Markers associated with HFN under two models for structure. Values are FDR (q-value) following both structured association and genomic control.

Marker	Chromosome	cM	*k*6: MAF 0.1	*k*11: MAF 0.1
wPt-0708	Unassigned	Unmapped	0.991093	0.223178
wPt-8157	Unassigned	Unmapped	0.991093	0.229754
wPt-2654	Unassigned	Unmapped	0.997308	0.238927
wPt-4652	1A|1B	Unmapped	0.991093	0.238927
wPt-7281	1A|7A	Unmapped	0.991093	0.238927
wPt-5312	1B	11.93525	0.991093	0.238927
wPt-1176	1B	Unmapped	0.991093	0.212316
wPt-4343	1B	Unmapped	0.991093	0.229754
wPt-10259	1B	Unmapped	0.991093	0.238927
wPt-0170	1B	Unmapped	0.991093	0.238927
wPt-1997	1B	Unmapped	0.991093	0.238927
wPt-5740	1B	Unmapped	0.991093	0.199317
wPt-1613	1B	Unmapped	0.991093	0.199317
wPt-5281	1B	Unmapped	0.991093	0.199317
wPt-7340	3A|3B	Unmapped	0.991093	0.199317
wPt-8355	3B	Unmapped	0.991093	0.199317
wPt-5003	4A	77.83858	0.997308	0.238927
wPt-1241	6B	35.78124	0.991093	0.238927
wPt-0884	7B	84.28452	0.991093	0.238927
wPt-9813	7B	84.55005	0.991093	0.199317
wPt-6320	7B	84.59786	0.991093	0.199317
wPt-1533	7B	85.23642	0.991093	0.199317
wPt-0126	7B	Unmapped	0.991093	0.199317
wPt-2449	7B	Unmapped	0.991093	0.238927
wPt-3785	7B	Unmapped	0.991093	0.199317
wPt-4644	7B	Unmapped	0.991093	0.199317
wPt-4319	7B|7D	Unmapped	0.991093	0.199317

Table 7-8. Markers associated with yield under two models for structure. Values are FDR (q-value) following both structured association and genomic control.

Marker	Chromosome	cM	*k*6: MAF 0.1	*k*11: MAF 0.1
wPt-8049	Unassigned	Unmapped	0.026985	0.202428
wPt-3467	Unassigned	Unmapped	0.026985	0.000488
wPt-9839	Unassigned	Unmapped	0.051065	0.210199
wPt-8328	2A	76.42814	0.045418	0.210199
wPt-3611	2A	76.52926	0.051065	0.210199
wPt-5027	2A	76.65359	0.026985	0.171452
wPt-1368	2A	76.69974	0.026985	0.171452
wPt-6148	2A	76.96541	0.041182	0.171452
wPt-9624	2A	77.81613	0.045586	0.24413
wPt-3976	2A	78.04085	0.046043	0.24413
wPt-6207	2A	78.09212	0.051065	0.210199
wPt-7626	2A	78.4962	0.043759	0.232989
wPt-8242	2A	Unmapped	0.026985	0.202428
rPt-6244	2A	Unmapped	0.026985	0.171452
wPt-8925	2A	Unmapped	0.026985	0.171452
wPt-5839	2A	Unmapped	0.026985	0.210199
wPt-1591	2A	Unmapped	0.041182	0.22345
wPt-0102	2A	Unmapped	0.041182	0.198832
wPt-6205	2A	Unmapped	0.043788	0.210199
wPt-9712	2A	Unmapped	0.045418	0.210199
wPt-6431	2A	Unmapped	0.045418	0.210199
wPt-8464	2A	Unmapped	0.051065	0.335063
wPt-2309	2A	Unmapped	0.051065	0.244938
wPt-3653	2A	Unmapped	0.06221	0.210199
wPt-1112	2A	Unmapped	0.069465	0.24413
wPt-7175	2A	Unmapped	0.070624	0.208016
wPt-2689	3D	89.55122	0.051065	0.210199
wPt-4255	6A	30.25088	0.026985	0.210199
wPt-4720	6B	12.51135	0.041182	0.171452
wPt-7150	6B	12.74009	0.049157	0.210199
wPt-4867	6B	12.81044	0.075993	0.210199
wPt-1922	6B	12.85368	0.041182	0.171452
wPt-3130	6B	12.85368	0.041182	0.171452
wPt-4386	6B|7A	Unmapped	0.026985	0.171452
wPt-9015	6B|7A	Unmapped	0.041182	0.171452
wPt-6563	6B|7A	Unmapped	0.041182	0.171452
wPt-5383	6B|7A	Unmapped	0.049918	0.198832
wPt-4678	6B|7A	Unmapped	0.051065	0.210199
wPt-0882	6B|7A	Unmapped	0.06221	0.210199
wPt-0452	6B|7A	Unmapped	0.066455	0.210199

Specific weight and **protein** do not yield any significant discoveries if genomic control is applied after structured association and the P-P plots suggest there is residual genome wide

inflation which should be controlled. It could therefore be argued that this method fails to discover any significant associations for these traits. For later comparison, however, I have identified significant markers discovered after structured association in Table 7-9 and Table 7-10. In the case of specific weight, seventeen markers are identified with FDR < 0.1 (*k*=6) and four additional with FDR < 0.1 (K=11) together these appear to suggest 6QTL. Table 7-10 shows the to twenty markers most significantly associated with protein (*k*=6) together with the twenty most associated (*k*=11) and duplicates removed. The result seems only to confirm that in this case inflation of the test statistic is such that there are an excess of discoveries.

Table 7-9. Markers associated with specific weight under two models for structure. Values are FDR (q-value) following both structured association and genomic control.

Marker	Chromosome	cM	*k*6: MAF 0.1	*k*11: MAF 0.1
wPt-6047	3B	22.20562	0.056359	0.116195
wPt-0773	3B	Unmapped	0.017665	0.052135
wPt-11407	3B	Unmapped	0.056359	0.104769
wPt-6239	3B	Unmapped	0.056879	0.104769
wPt-2291	4A	80.86056	0.056359	0.104769
wPt-9824	4B	2.743866	0.056879	0.041817
wPt-2707	5B	65.06206	0.056879	0.126473
wPt-7238	5B	86.63695	0.177999	0.061958
wPt-7006	5B	Unmapped	0.056359	0.061958
wPt-7059	5B	Unmapped	0.098986	0.126473
wPt-3115	5B	Unmapped	0.177999	0.061958
wPt-7475	6A	26.55221	0.104655	0.071429
wPt-4255	6A	30.25088	0.079922	0.229181
wPt-9832	6A	Unmapped	0.073175	0.126473
wPt-1241	6B	35.78124	0.056879	0.117572
wPt-3733	6B	60.42408	0.056359	0.041817
wPt-7049	Unassigned	Unmapped	0.056393	0.126915
wPt-6191	Unassigned	Unmapped	0.056879	0.126473
wPt-3467	Unassigned	Unmapped	0.056879	0.311937
wPt-8049	Unassigned	Unmapped	0.079922	0.236599

Table 7-10. Markers associated with protein under two models for structure. Values are FDR (q-value) following both structured association and genomic control.

Marker	Chromosome	cM	*k*6: MAF 0.1	*k*11: MAF 0.1
wPt-0196	1A	Unmapped	0.066982	0.099048
wPt-6012	1B	11.71326	0.052591	0.099048
wPt-1973	1B	Unmapped	0.07375	0.138211
wPt-1573	1B	Unmapped	0.109252	0.144222
wPt-7212	3B	13.31478	0.017952	0.099048
wPt-8355	3B	Unmapped	0.02108	0.15944
wPt-3397	3B	Unmapped	0.022184	0.282201
wPt-10201	3B	Unmapped	0.023235	0.251384
wPt-4608	3B	Unmapped	0.09336	0.138211
wPt-11407	3B	Unmapped	0.039121	0.144222
wPt-2689	3D	89.55122	0.014025	0.121757
wPt-0023	4A	Unmapped	0.018662	0.144222
rPt-2478	4A|7A	Unmapped	0.25905	0.121757
rPt-7987	4A|7A	Unmapped	0.271074	0.144222
wPt-1964	4D	Unmapped	0.008124	0.121757
tPt-4627	4D	Unmapped	0.017952	0.153926
wPt-3058	4D	Unmapped	0.039121	0.144222
wPt-2707	5B	65.06206	0.020996	0.273161
wPt-9692	6A	Unmapped	0.315739	0.121757
wPt-7906	6A	Unmapped	0.227892	0.138211
wPt-7150	6B	12.74009	0.020975	0.176238
wPt-3304	6B	12.77263	0.020996	0.161242
wPt-4867	6B	12.81044	0.020996	0.153926
wPt-4678	6B|7A	Unmapped	0.017952	0.144222
wPt-0882	6B|7A	Unmapped	0.020975	0.153926
wPt-6585	6B|7A	Unmapped	0.02108	0.294651
wPt-0687	7A	Unmapped	0.020996	0.418833
wPt-6967	7A	Unmapped	0.266356	0.132056
wPt-4555	7D	76.92254	0.006785	0.132056
wPt-0231	7D	Unmapped	0.008124	0.144222
wPt-6191	Unassigned	Unmapped	0.018662	0.233376
wPt-6417	Unassigned	Unmapped	0.018662	0.121757
wPt-4947	Unassigned	Unmapped	0.039121	0.132056
wPt-5479	Unassigned	Unmapped	0.299384	0.132056

7.4 Correction using principal coordinates

Using eigenvectors of a covariance matrix as a model for structure and applying the statistical correction for structure proposed by Price *et al.* (2006) had the effect of reducing the inflation of the test statistic. This was most notable for the phenotypes which had already been demonstrated to be most strongly associated with population substructure (yield and protein).

The choice of the number of principal coordinates is at the discretion of the analyst and it seems likely (7.1.2) that the suitable value will be in the range 10 to 14 PCs depending on phenotype. Analysis was conducted for 10, 12 and 14 vectors for protein and yield and 10, 12, 14 and 16 vectors for HFN and specific weight (the additional vector being added because deflation was incomplete with 14 PCs). The results are summarised as PP plots in Figure 6-25 to Figure 7-27. As discussed in the context of Structured association, it is possible to use GC as an additional treatment if the deflation by the regression model is incomplete, the effect of this is also presented in the figures.

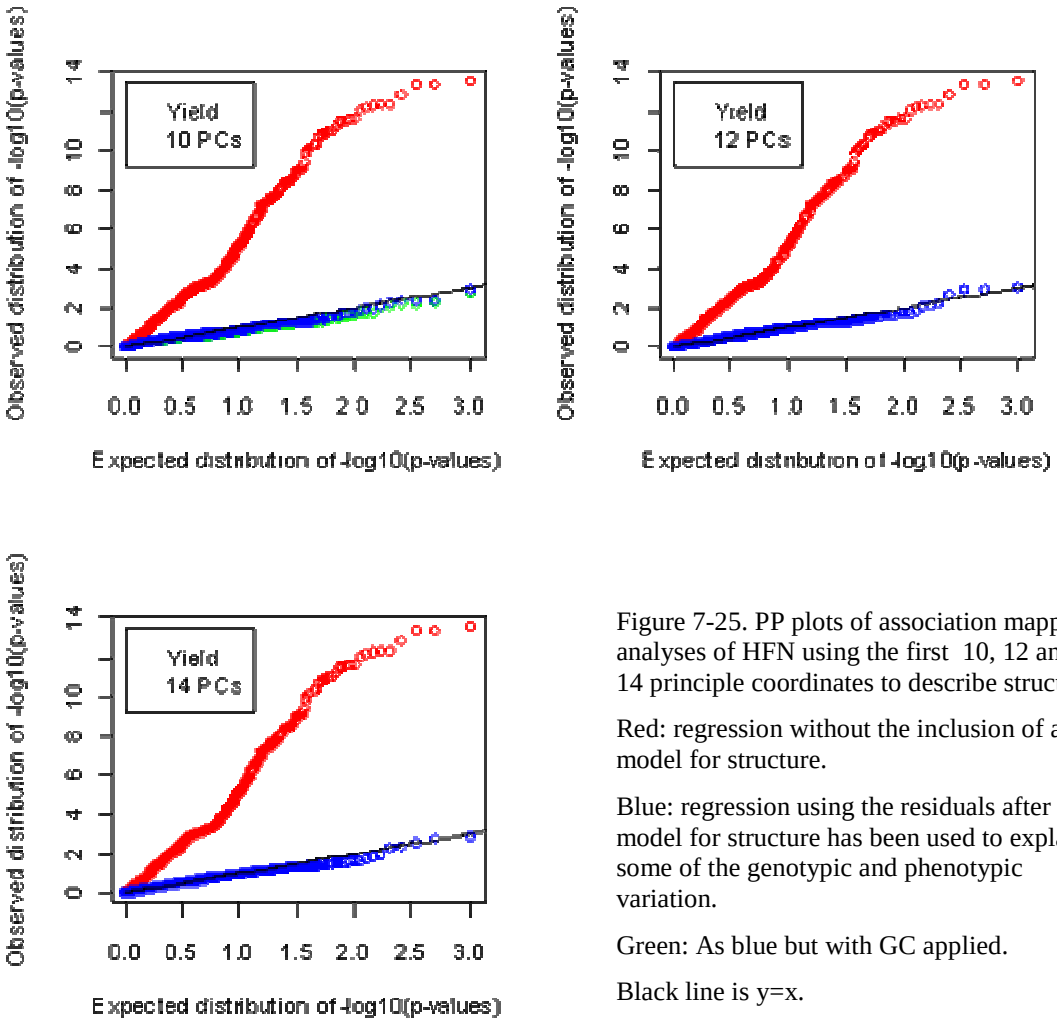


Figure 7-25. PP plots of association mapping analyses of HFN using the first 10, 12 and 14 principle coordinates to describe structure.

Red: regression without the inclusion of a model for structure.

Blue: regression using the residuals after the model for structure has been used to explain some of the genotypic and phenotypic variation.

Green: As blue but with GC applied.

Black line is y=x.

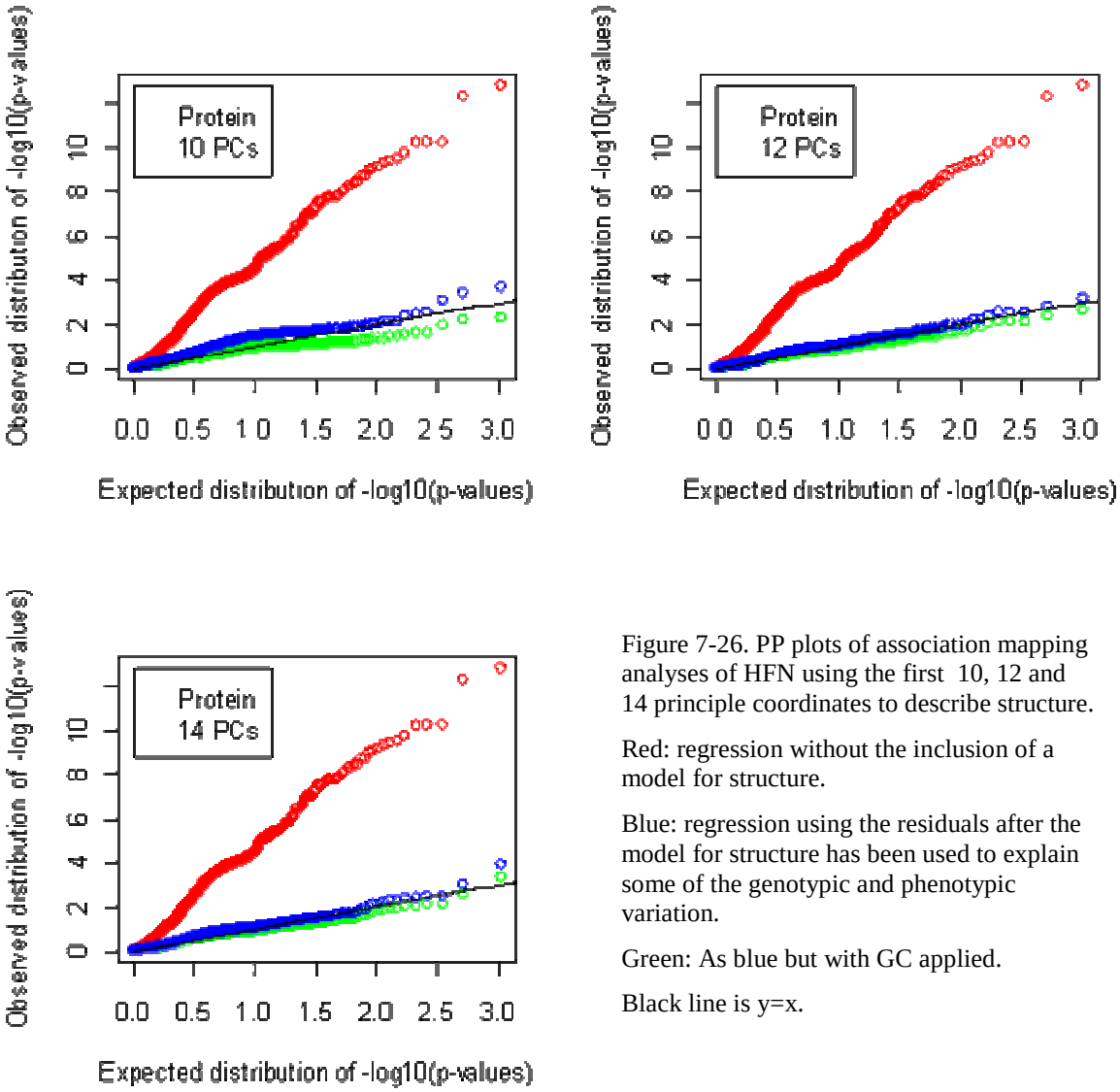


Figure 7-26. PP plots of association mapping analyses of HFN using the first 10, 12 and 14 principle coordinates to describe structure.

Red: regression without the inclusion of a model for structure.

Blue: regression using the residuals after the model for structure has been used to explain some of the genotypic and phenotypic variation.

Green: As blue but with GC applied.

Black line is y=x.

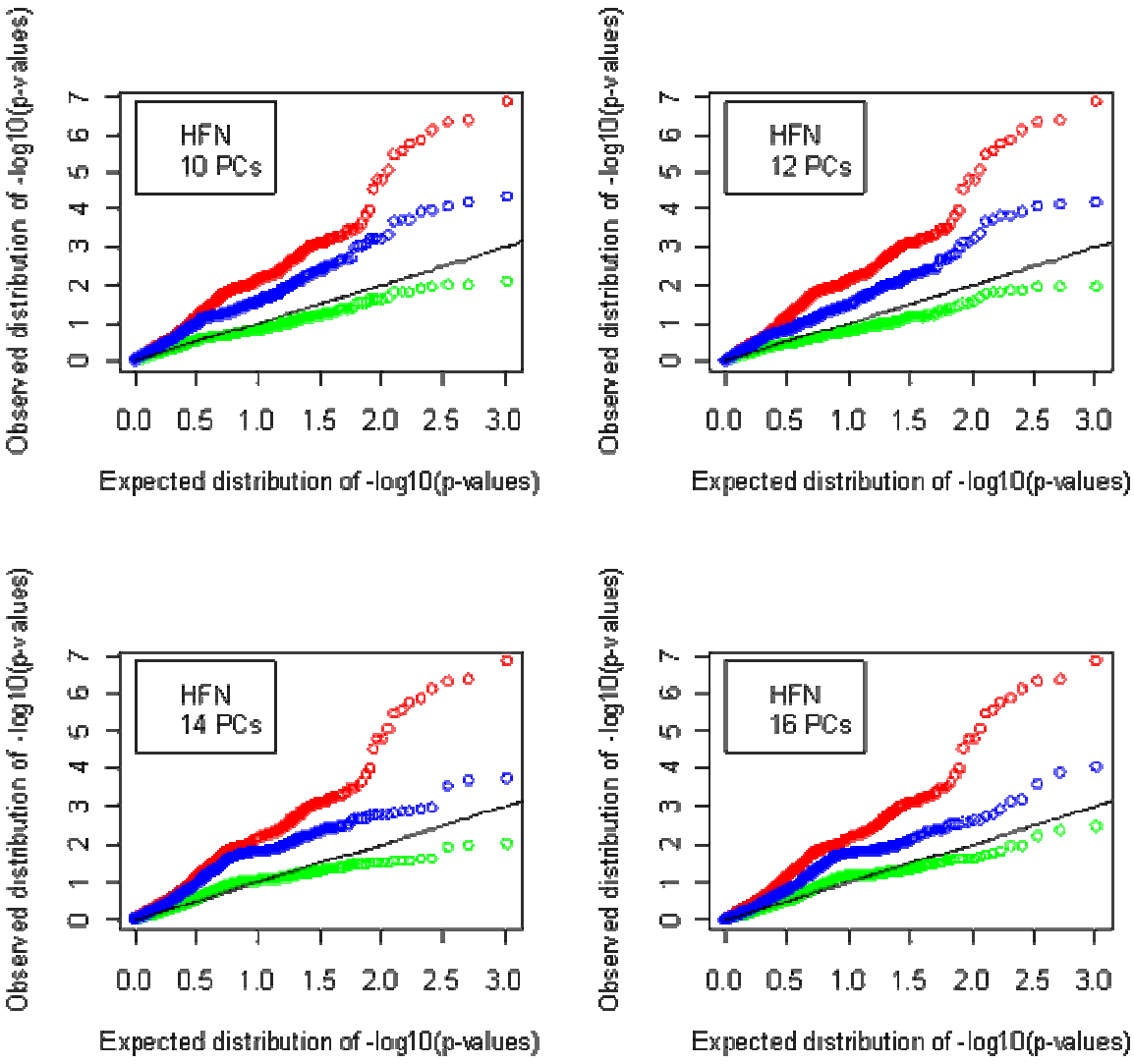


Figure 7-27. PP plots of association mapping analyses of HFN using the first 10, 12 ,14 and 16 principal coordinates to describe structure. Red: regression without the inclusion of a model for structure. Blue: regression using the residuals after the model for structure has been used to explain some of the genotypic and phenotypic variation. Green: As blue but with GC applied. Black line is y=x.

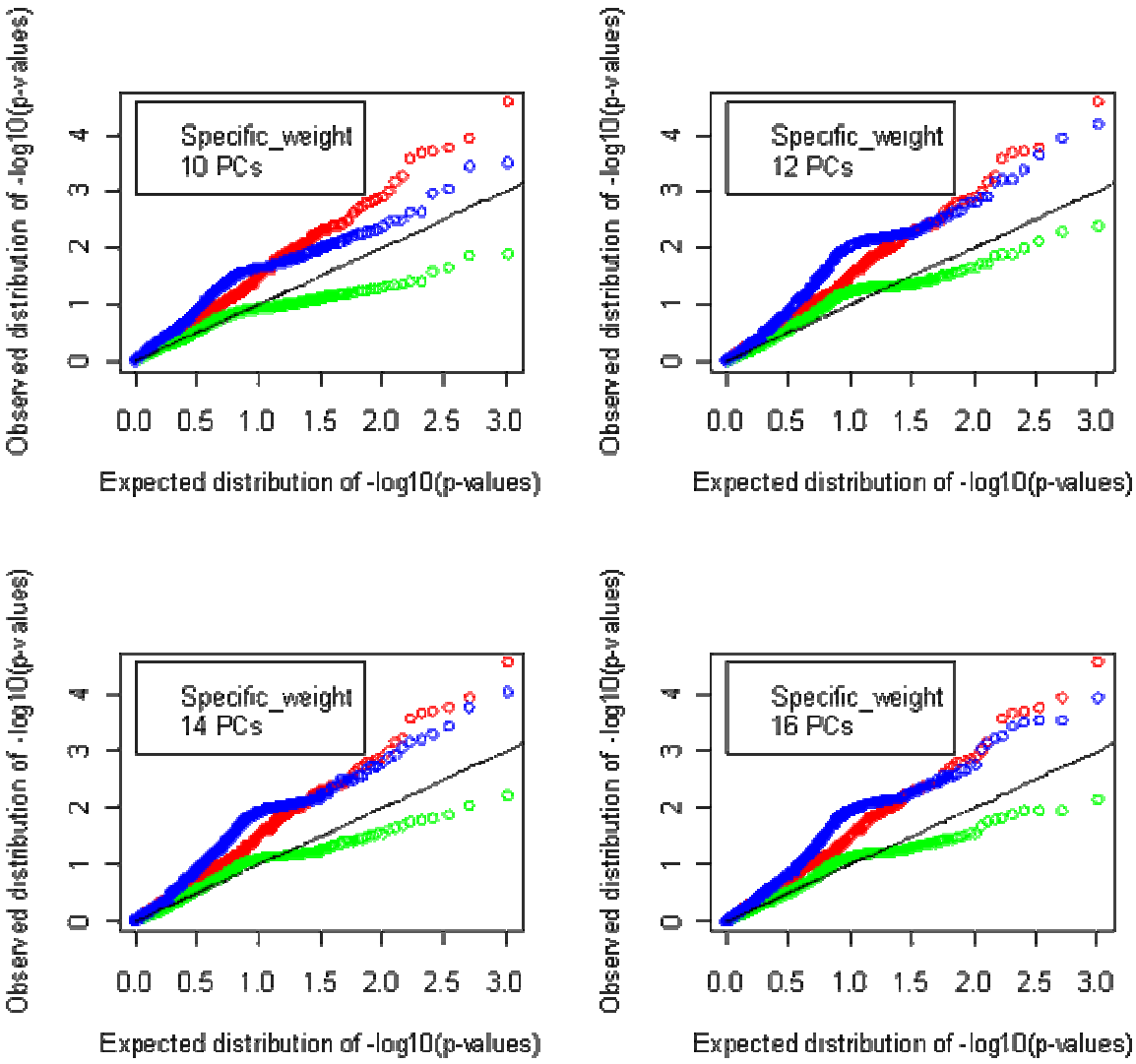


Figure 7-28. PP plots of association mapping analyses of specific weight using the first 10, 12, 14 and 16 principal coordinates to describe structure. Red: regression without the inclusion of a model for structure. Blue: regression using the residuals after the model for structure has been used to explain some of the genotypic and phenotypic variation. Green: As blue but with GC applied. Black line is y=x.

7.4.1 QTL discovery using the PC method.

The P-P plots show only slight evidence for statistically significant markers for any of the traits. However, the P-P plots for HFN and specific weight have been distorted by the statistical correction and are no longer linear, at the lowest p-values for both there appears to be a deflection upward from the trend in the last few tens of markers which might prove to be

statistically significant under a different model. This is most noticeable for specific weight with 16 PCs and HFN with 14 PCs (Figure 7-27 and Figure 7-28).

Calculating q-values corresponding to the observed distribution of p-values from these analyses allows an estimate of the false discovery rate for each marker (Figure 7-29 to Figure 7-32), as had been discussed this is a useful alternative to setting a significance threshold in terms only of p-value. Using only 10 PCs to define structure returns no 'low FDR' markers for any phenotype. As the number of PCs rise the number of significant markers fall – as the confounding effect of structure is removed. There remain (weakly) significant markers for yield (12 and 14 PCs) and the single marker for protein (14 and 16 PCs). Table 7-11 and Table 7-12 identify the ten most significant (in terms of p-value) markers for yield and protein and the corresponding FDR values. It is notable that these tables do not contain any of the ten most significant markers identified following naïve regression analysis (Table 7-3 and Table 7-6).

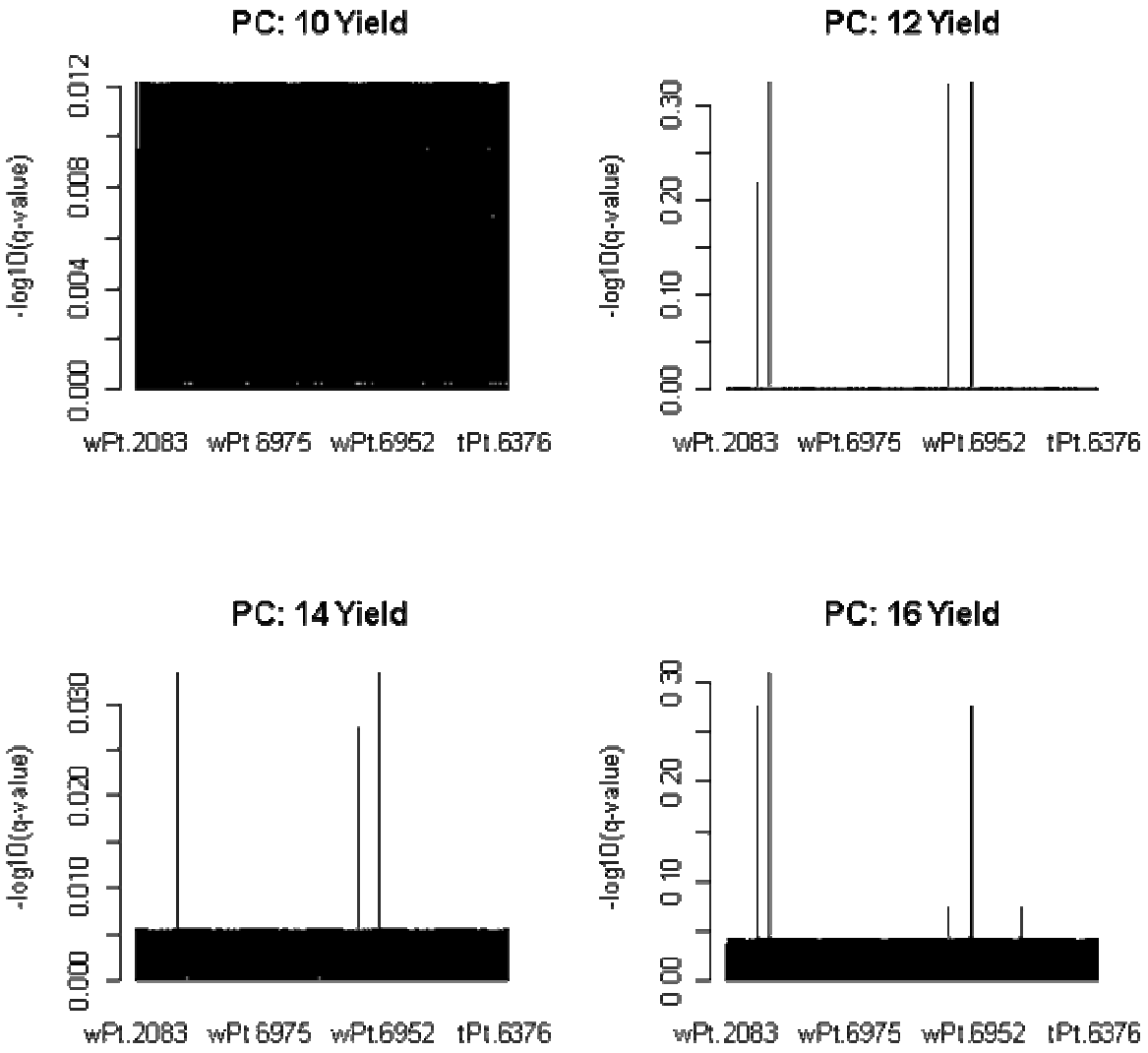


Figure 7-29. Association between marker genotype and yield; the indicated number of principal coordinates have been used to model structure and residual inflation of the test statistic has been removed by genomic control. There is no association signal when 10 PCs are used in the model but four markers show a weak (FDR ≈ 50%) when the number of PCs is 12 or 16 and three of the markers are identified by both models.

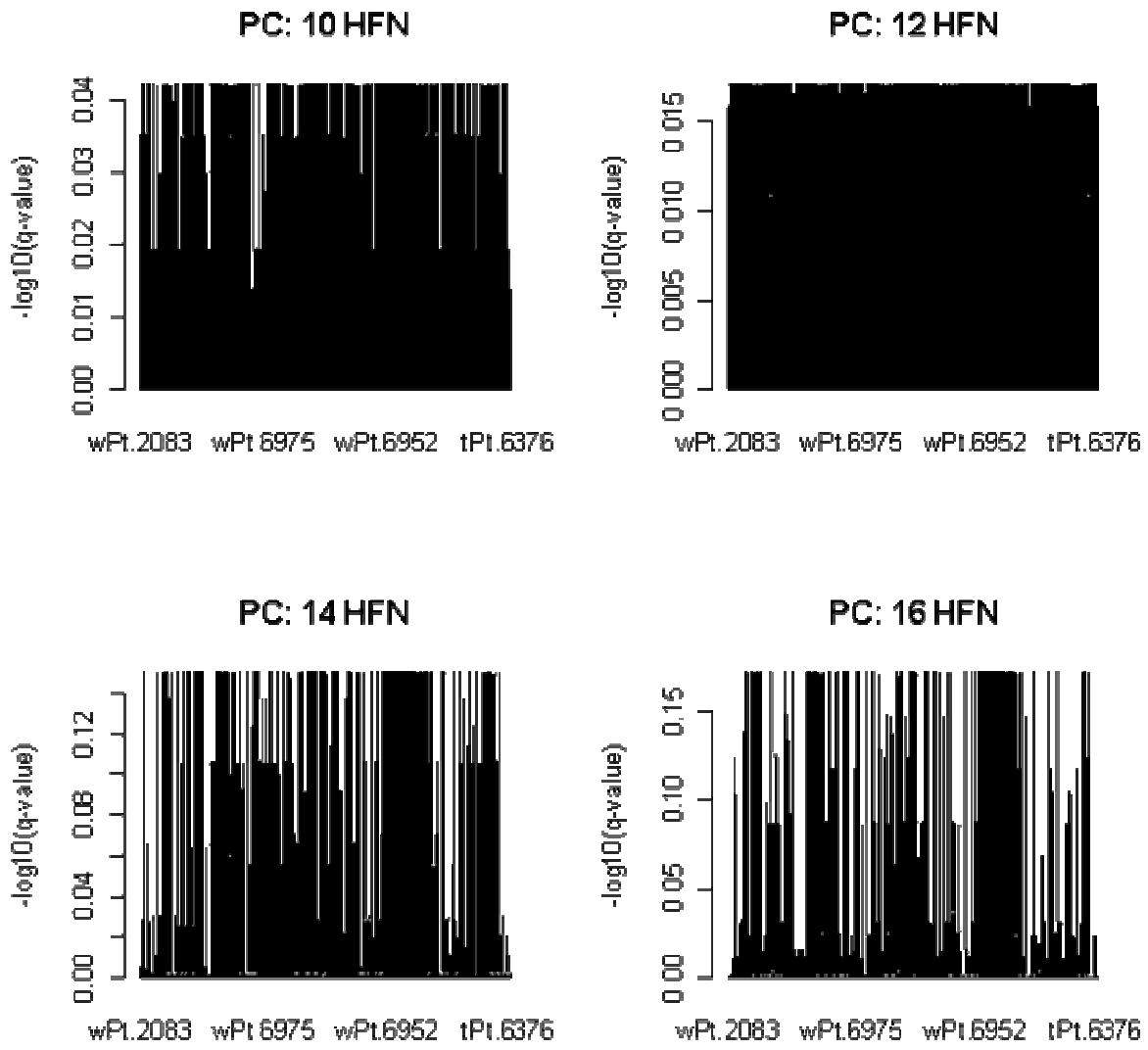


Figure 7-30. Association between marker genotype and HFN; the indicated number of principal coordinates have been used to model structure and residual inflation of the test statistic has been removed by genomic control. There is no association signal when 10 or 12 PCs are used in the model. Weakly associated clusters of markers appear when 14 and 16 PCs are used but the genome-wide nature of their distribution suggests they are artifactual.

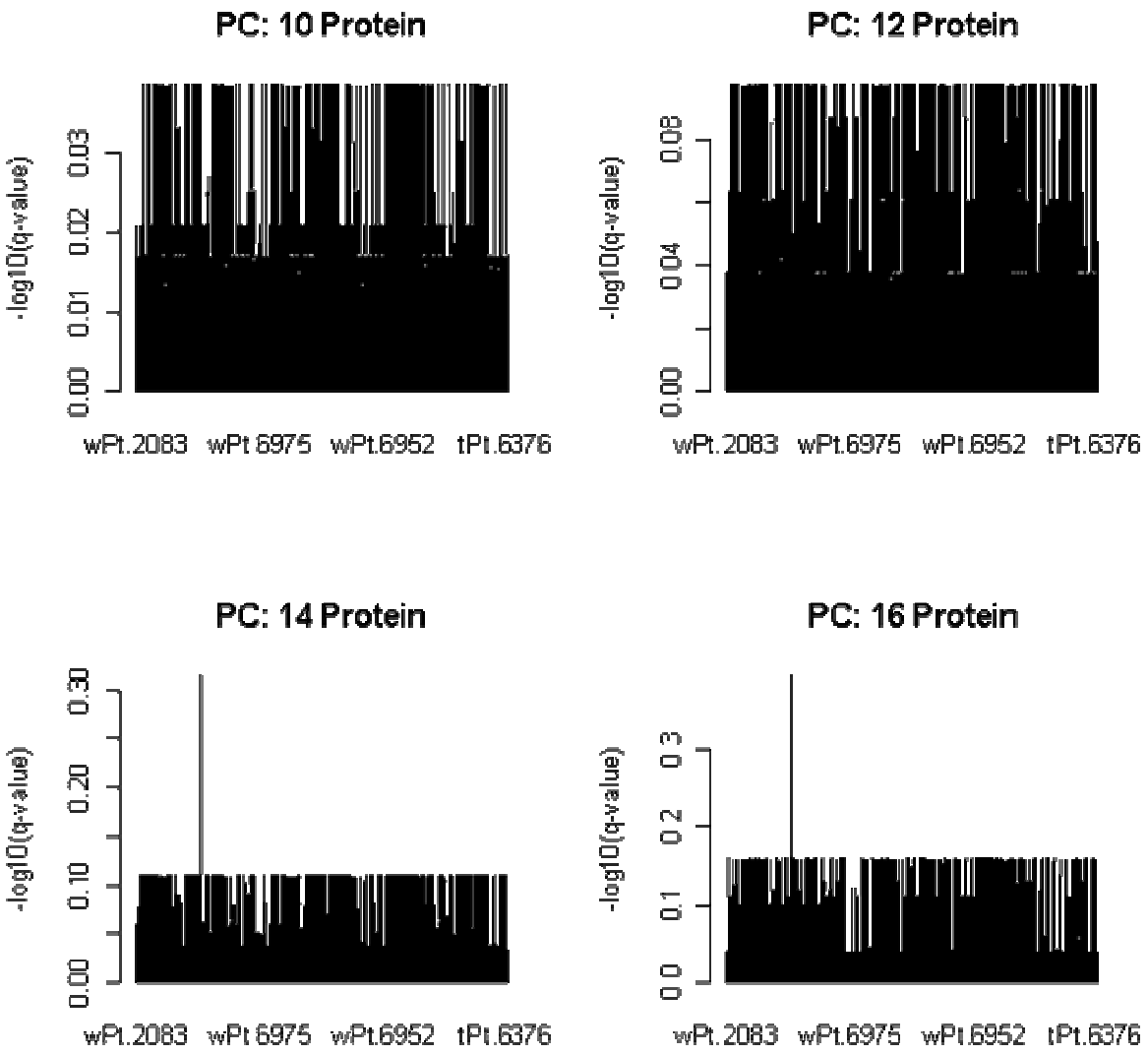


Figure 7-31. Association between marker genotype and protein; the indicated number of principal coordinates have been used to model structure and residual inflation of the test statistic has been removed by genomic control. There is no association signal when 10 or 12 PCs are used in the model but one marker shows a weak (FDR ≈ 40%) association when the number of PCs is 14 or 16.

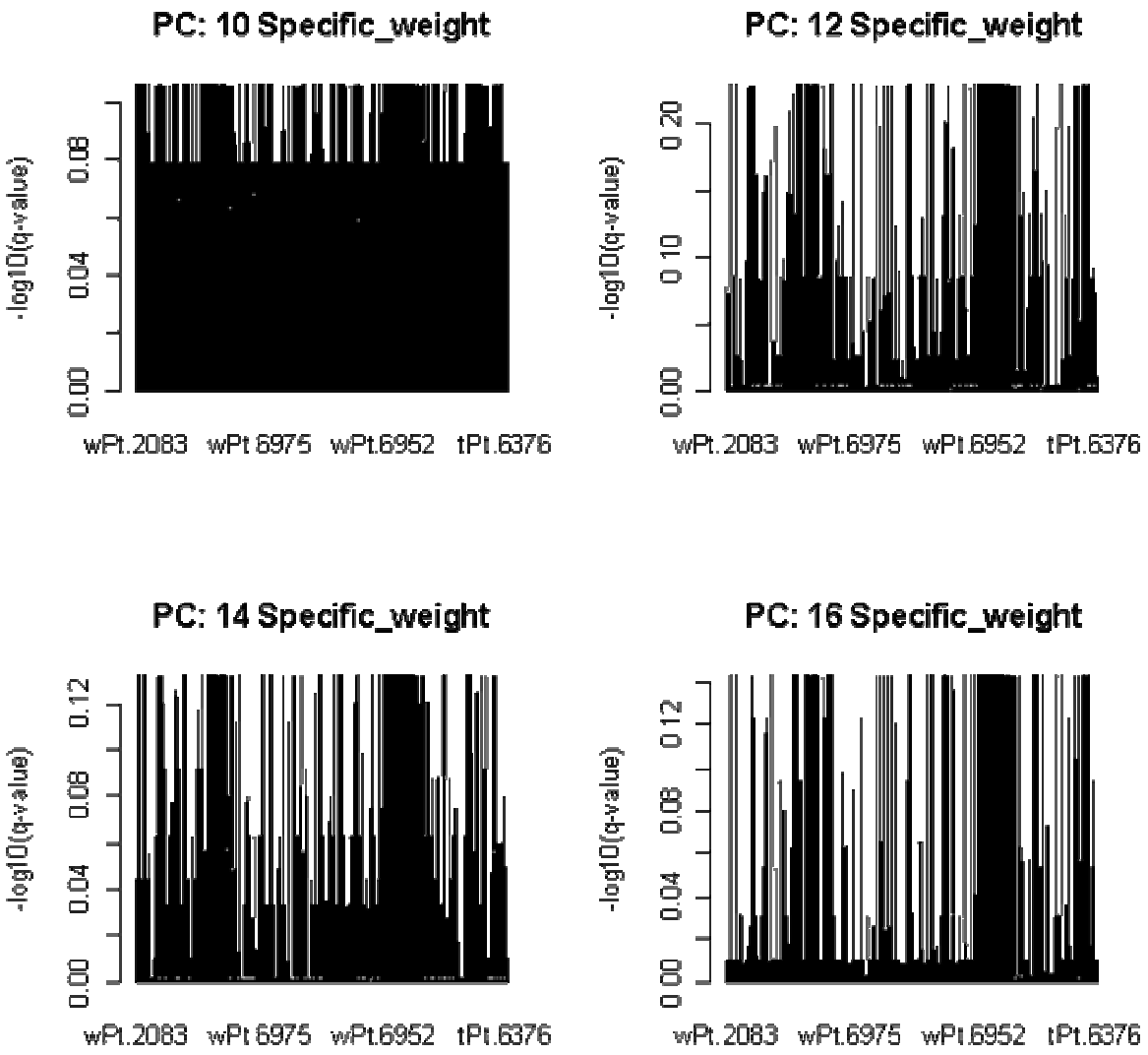


Figure 7-32. Association between marker genotype and specific weight; the indicated number of principal coordinates have been used to model structure and residual inflation of the test statistic has been removed by genomic control. There is no association signal when 10 PCs are used in the model with greater numbers of PCs there are many weak association signals but their genome-wide nature suggests they may be artefactual.

Table 7-11 FDR for ten markers most associated with yield following statistical correction using 12 principal coordinates as a model for structure.

Marker	Chromosome	cM	p-value following Eigenstrat	FDR following Eigenstrat + GC
wPt-5801	1B	12.55688	0.015628	>0.9
wPt-0725	1B	unmapped	0.00937	>0.9
wPt-8548	2B	unmapped	0.011992	>0.9
wPt-7612	4A	unmapped	0.00437	>0.9
wPt-1272	4B	6.459287	0.014964	>0.9
wPt-3439	5B	3.165375	0.003939	0.47
wPt-4553	7A	116.4446	0.000507	0.47
wPt-0961	7A	118.0137	0.001644	0.60
wPt-1269	7D	8.787941	0.001496	0.47
wPt-4280	unassigned	unmapped	0.013518	>0.9

Table 7-12 FDR for ten markers most associated with protein following statistical correction using 16 principal coordinates as a model for structure.

Marker	Chromosome	cM	p-value following Eigenstrat	FDR following Eigenstrat + GC
wPt-6417	unassigned	unmapped	0.001072	>0.9
wPt-1976	unassigned	unmapped	0.005203	>0.9
wPt-3403	unassigned	unmapped	0.006447	>0.9
wPt-0899	3A	unmapped	0.007235	>0.9
wPt-2559	3B	27.61748	0.012804	>0.9
wPt-1945	3B	unmapped	0.005999	>0.9
wPt-3342	3B	unmapped	0.013397	>0.9
wPt-8412	3B|6B	unmapped	0.000199	0.40
rPt-2478	4A|7A	unmapped	0.008975	>0.9
wPt-9832	6A	unmapped	0.003616	>0.9

7.5 Mixed effects modelling.

As with Structured Association and PC correction, the mixed model leaves the analyst with some discretion as to how, for example, the mathematical model of kinship is to be calculated. As with previous methods the purpose of the mathematical representation of structure/kinship is to predict phenotype prior to the addition of a marker data to the regression model. Intuitively then, the kinship model which best predicts phenotype will best control spurious association.

Genetic correlation (K_{gen}) and statistical correlation (K_{stat}) between standardised genotypes of varieties are both potential methods of estimating kinship (respectively; Astle and Balding 2010; Cockram *et al.* 2010) which take account of excess allele sharing. The key question is then the markers to be used to calculate the matrices. Clearly, there is an argument for maximising the number of markers since this ought to sample the genome most widely and thus measure differences in kinship most accurately. Conversely, it is inadvisable to simply use every available marker since groups of highly correlated markers may disproportionately represent small areas of the genome and give excessive weight to one aspect of the population's genetic history – for example the introgression of 1B1R into wheat – and so distort relationships. In Figure 7-33 to Figure 7-40 I present surface plots of the residual variance from a model in which phenotype is predicted in a mixed model with **K** as a random effect and no fixed effect of marker genotype. These surface plots suggest MAF should be less than 0.1 (consistent with the observation that most information about kinship is present in shared minor alleles) and maximum between marker R^2 appears to vary with trait (see Table 7-13).

Table 7-13 Values of MAF and maximum between marker R^2 which give the kinship matrices that minimise V_E for the wheat phenotypes.

Trait	Statistical Correlation		Genetic Correlation	
	MAF	R^2	MAF	R^2
Yield	0.05	0.7	0.05	0.9
HFN	0.07	0.8	0.05	1.0
Protein	0.05	0.8	0.05	0.6
Specific Weight	0.08	0.5	0.09	0.7

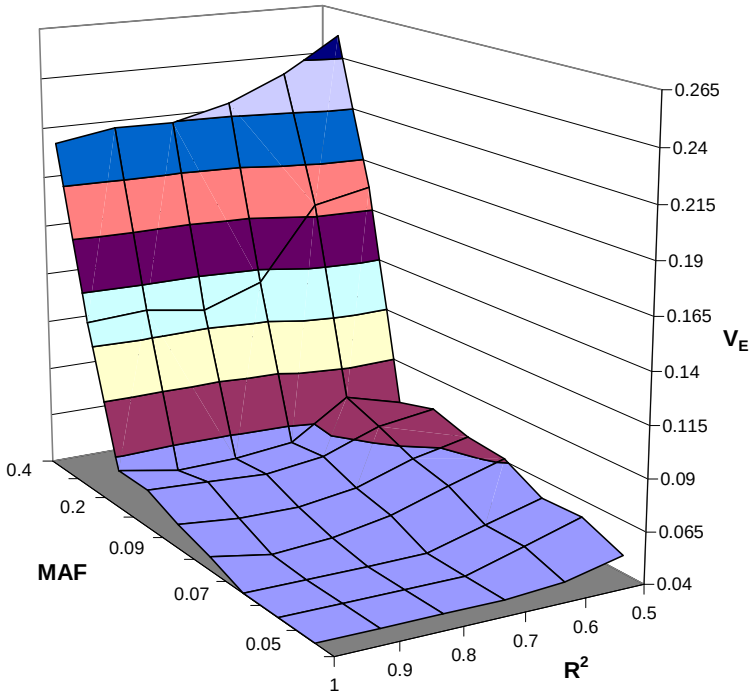


Figure 7-33Yield. Residual variance from a mixed model comprising only the random effect of kinship defined as genetic correlation based on standardised genotypes skimmed to indicated MAF and R2. Minimum Ve is 0.029 at MAF=0.05 and R^2 =0.9

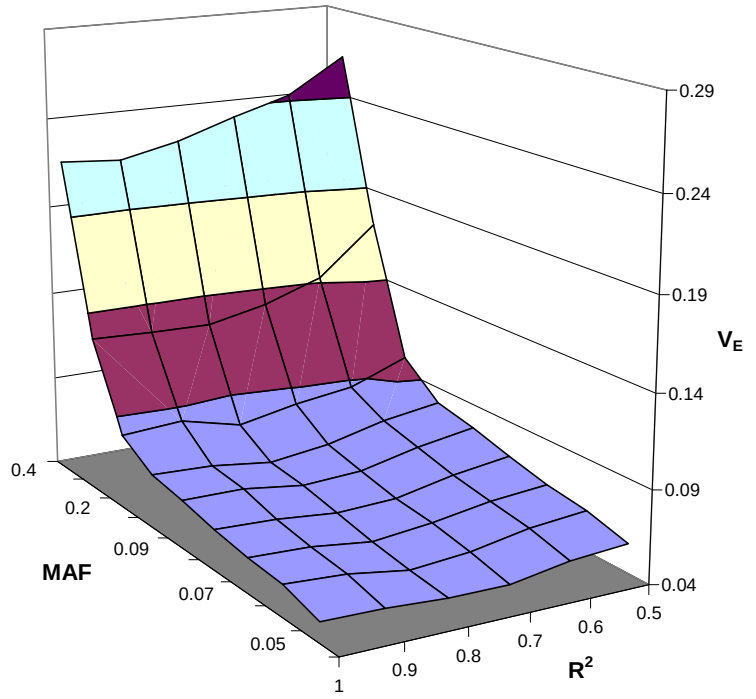


Figure 7-34. Yield. Residual variance from a mixed model comprising only the random effect of kinship defined as statistical correlation based on standardised genotypes skimmed to indicated MAF and R2. Minimum Ve is 0.048 at MAF=0.05 and R^2 =0.7

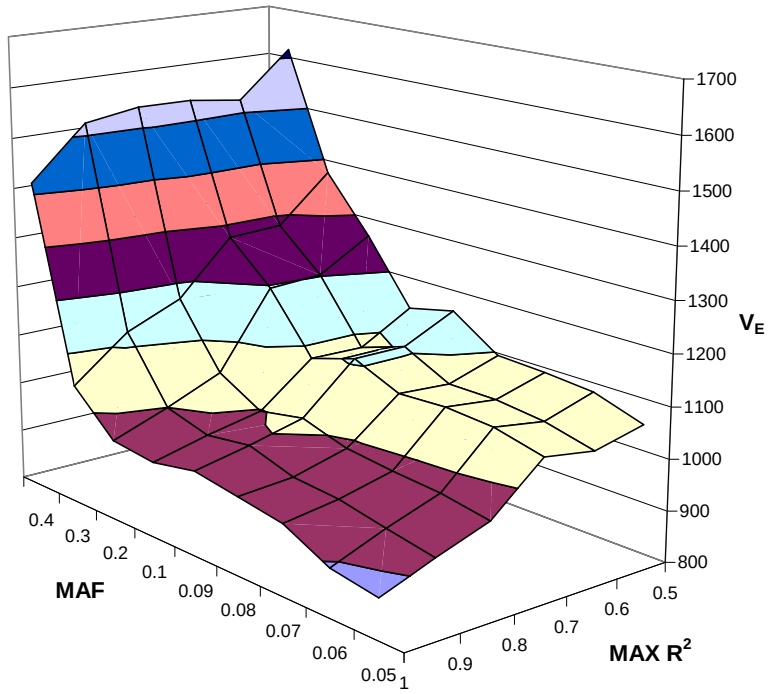


Figure 7-35. HFN. Residual variance from a mixed model comprising only the random effect of kinship defined as genetic correlation based on standardised genotypes skimmed to indicated MAF and R2. Minimum Ve is 878 at MAF=0.05 and R^2 =1

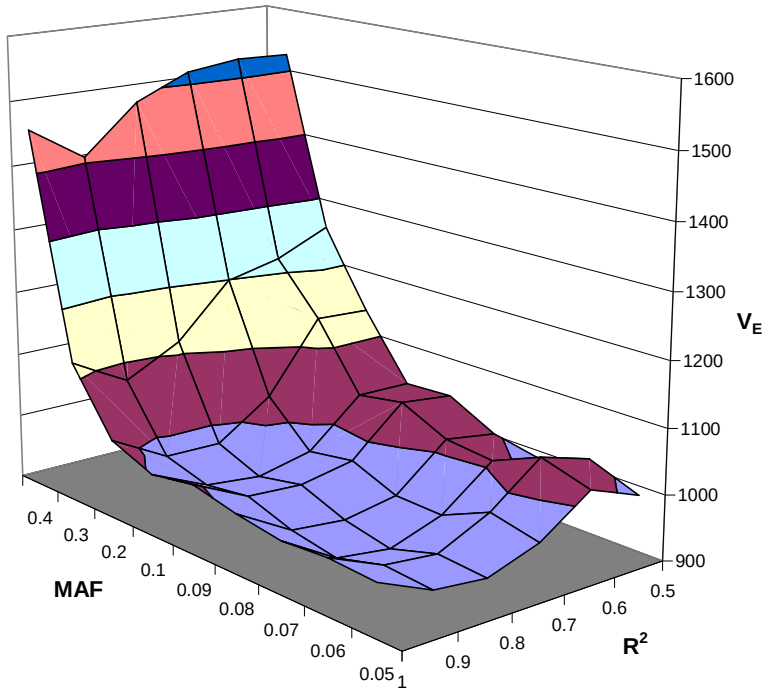


Figure 7-36. HFN. Residual variance from a mixed model comprising only the random effect of kinship defined as statistical correlation based on standardised genotypes skimmed to indicated MAF and R2. Min Ve is 919 at MAF=0.07 and R^2 =0.8

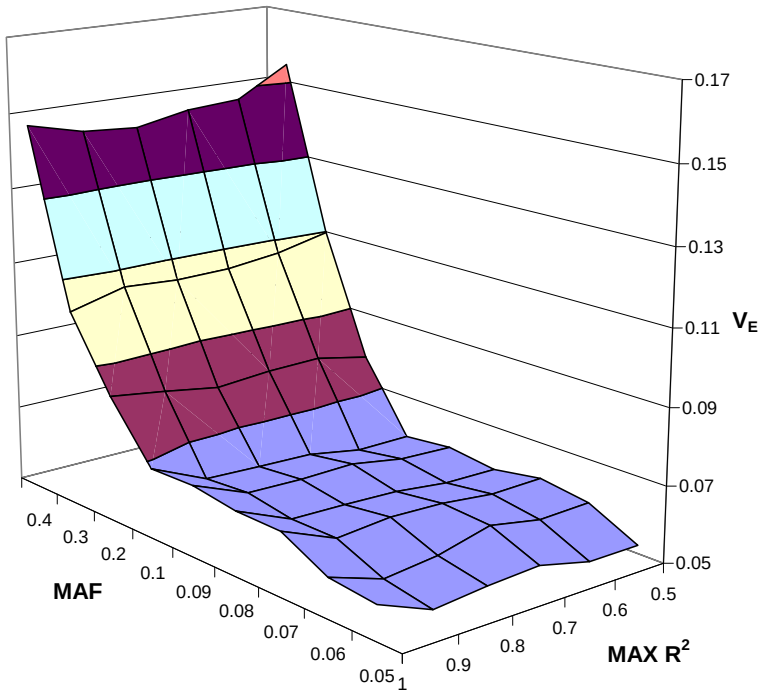


Figure 7-37. Protein. Residual variance from a mixed model comprising only the random effect of kinship defined as genetic correlation based on standardised genotypes skimmed to indicated MAF and R2. Minimum Ve is 0.05 at MAF=0.05 and R^2 =0.6

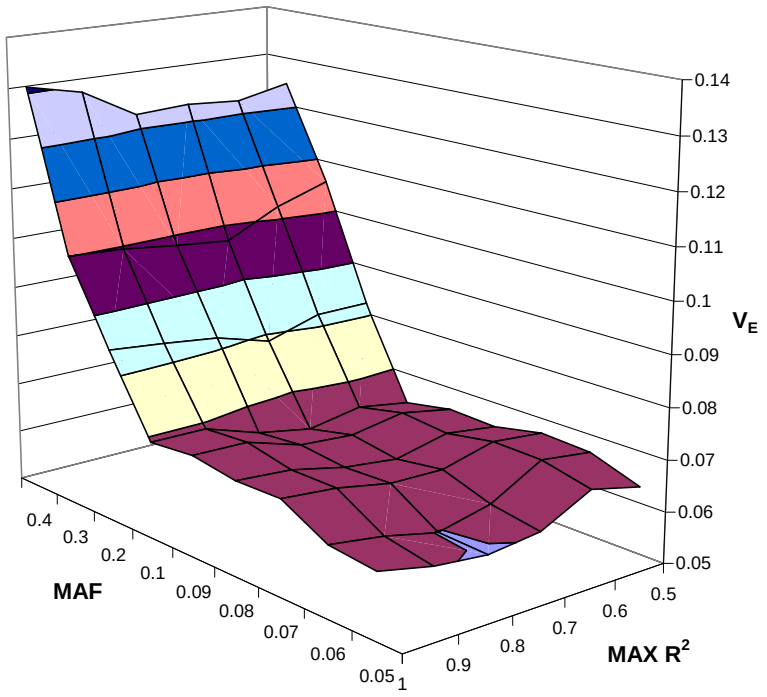


Figure 7-38. Protein. Residual variance from a mixed model comprising only the random effect of kinship defined as statistical correlation based on standardised genotypes skimmed to indicated MAF and R2. Minimum Ve is 0.059 at MAF=0.05 and R^2 =0.8

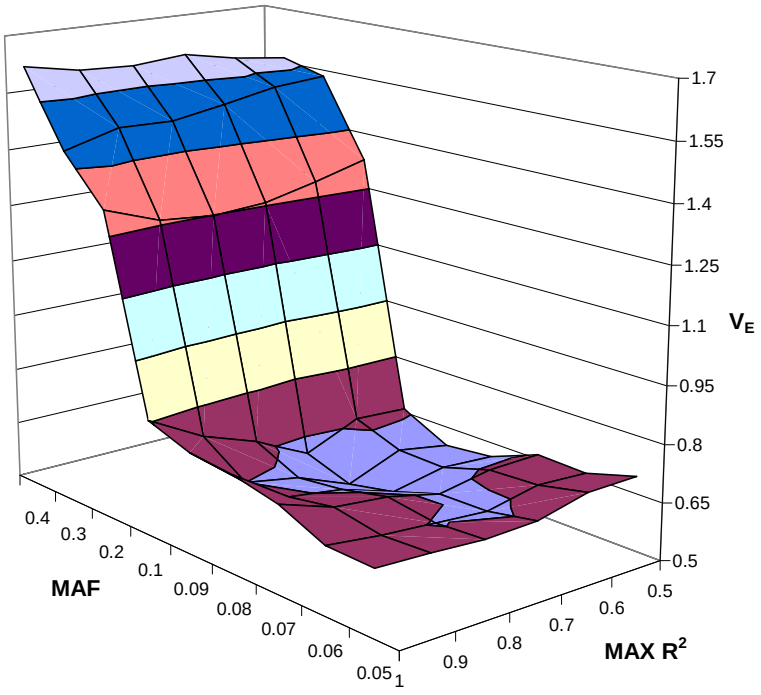


Figure 7-39. Specific weight. Residual variance from a mixed model comprising only the random effect of kinship defined as genetic correlation based on standardised genotypes skimmed to indicated MAF and R2. Minimum Ve is 0.57 at MAF=0.09 and R^2 =0.7.

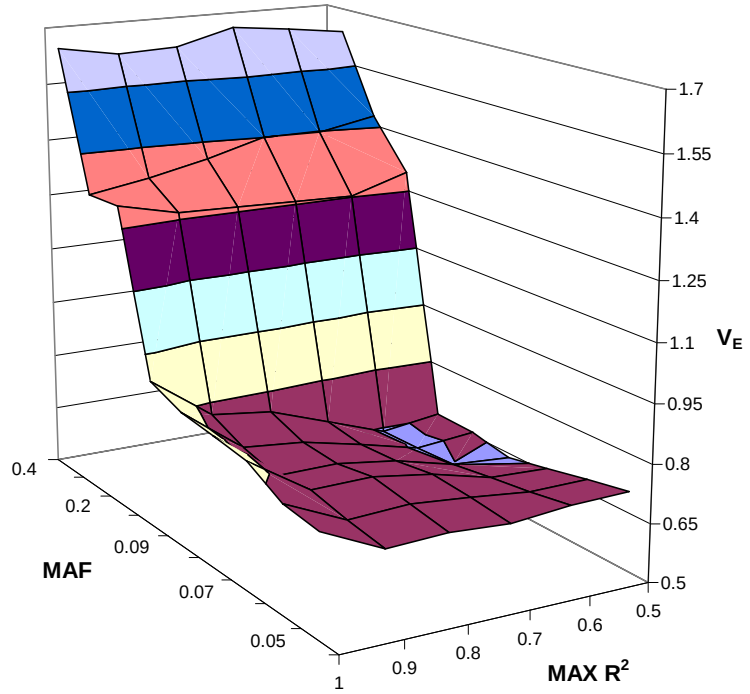


Figure 7-40. Specific weight. Residual variance from a mixed model comprising only the random effect of kinship defined as statistical correlation based on standardised genotypes skimmed to indicated MAF and R2. Min Ve is 0.64 at MAF=0.08 and R^2 =0.5.

Having established the optimum kinship models for each trait an association analysis may be performed using the optimum **K** matrices as random effects and markers as fixed effects. QQ plots of these analyses are presented in Figure 7-41 to Figure 7-44.

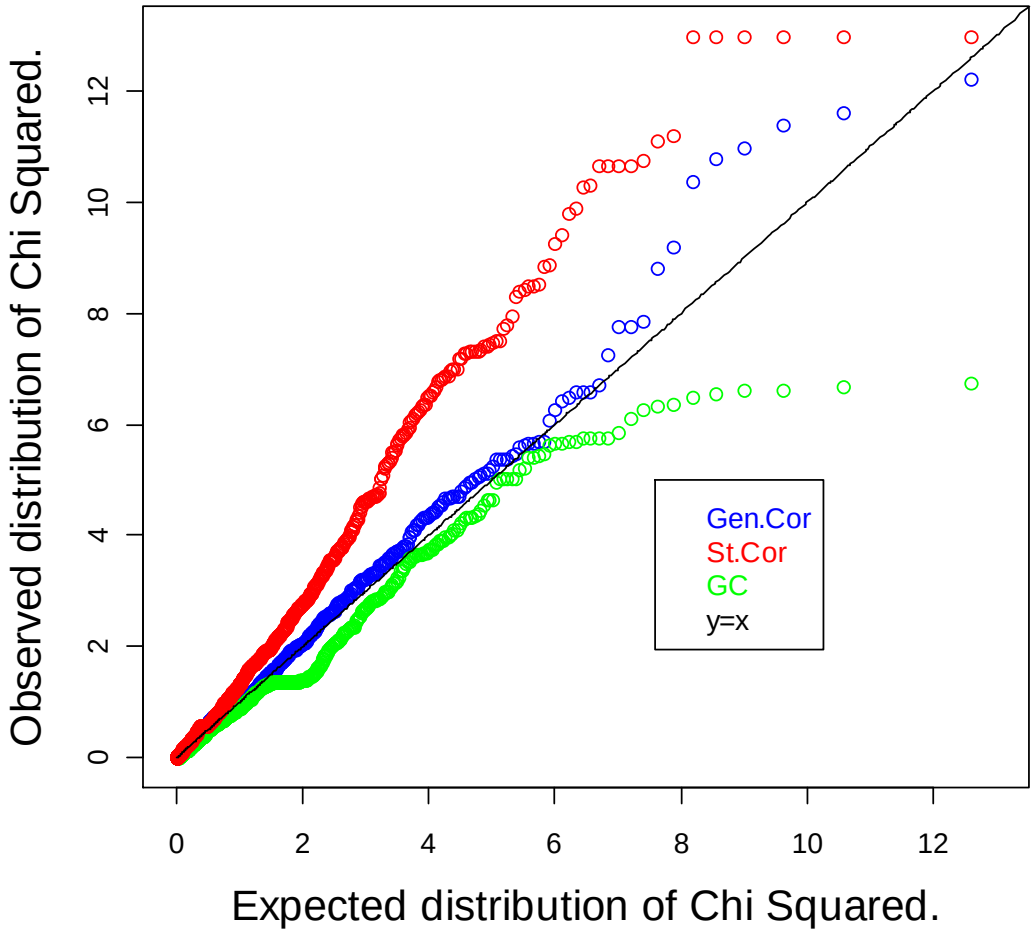


Figure 7-41 .Yield. QQ plot of Chi-Sq (1 d.f.) tests for association between marker and phenotype. Kinship has been modelled (MM) as statistical correlation (St.Cor) or genetic correlation (Gen.Cor) based on standardised genotype. MAF and maximum between marker R2 has been adjusted to the optimum values given in Table 7-13. Genomic control following an association test without adjustment for kinship is presented for comparison. Values greater than 13 have been set to 13.

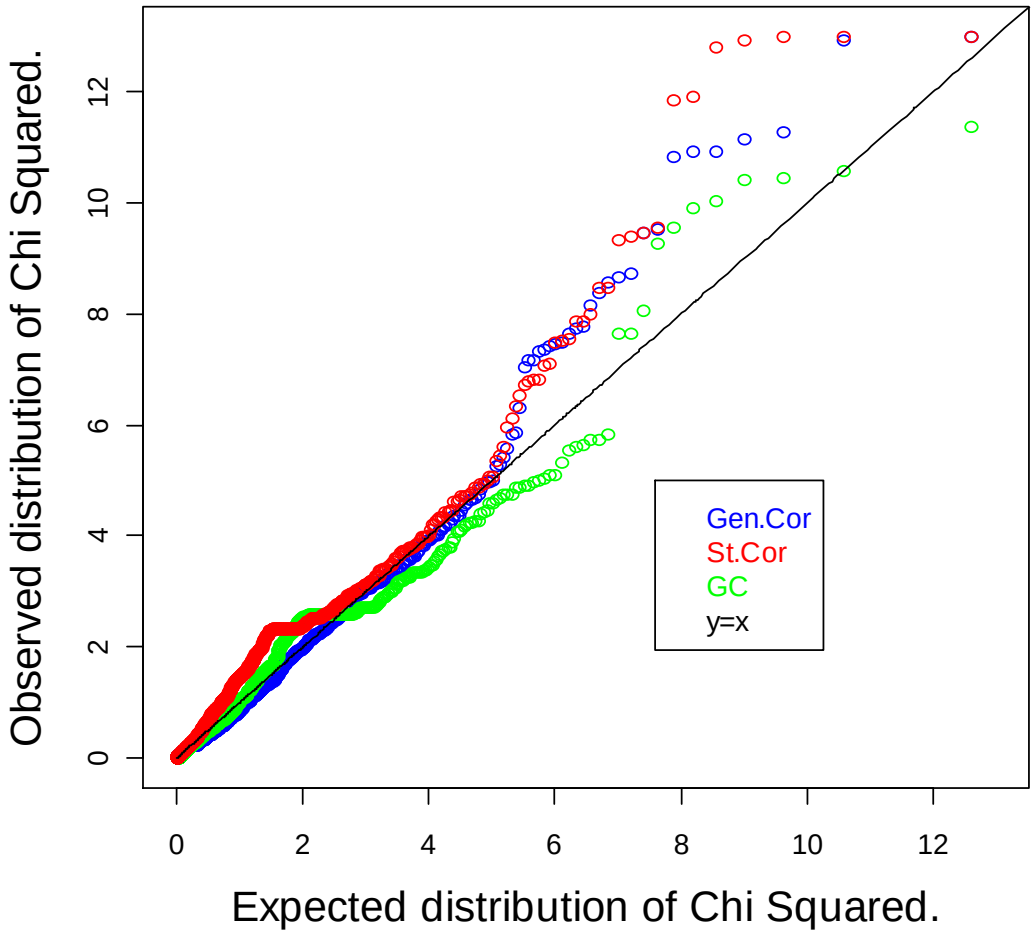


Figure 7-42 . HFN. QQ plot of Chi-Sq (1 d.f.) tests for association between marker and phenotype. Kinship has been modelled (MM) as statistical correlation (St.Cor) or genetic correlation (Gen.Cor) based on standardised genotype. MAF and maximum between marker R2 has been adjusted to the optimum values given in Table 7-13. Genomic control following an association test without adjustment for kinship is presented for comparison. Values greater than 13 have been set to 13.

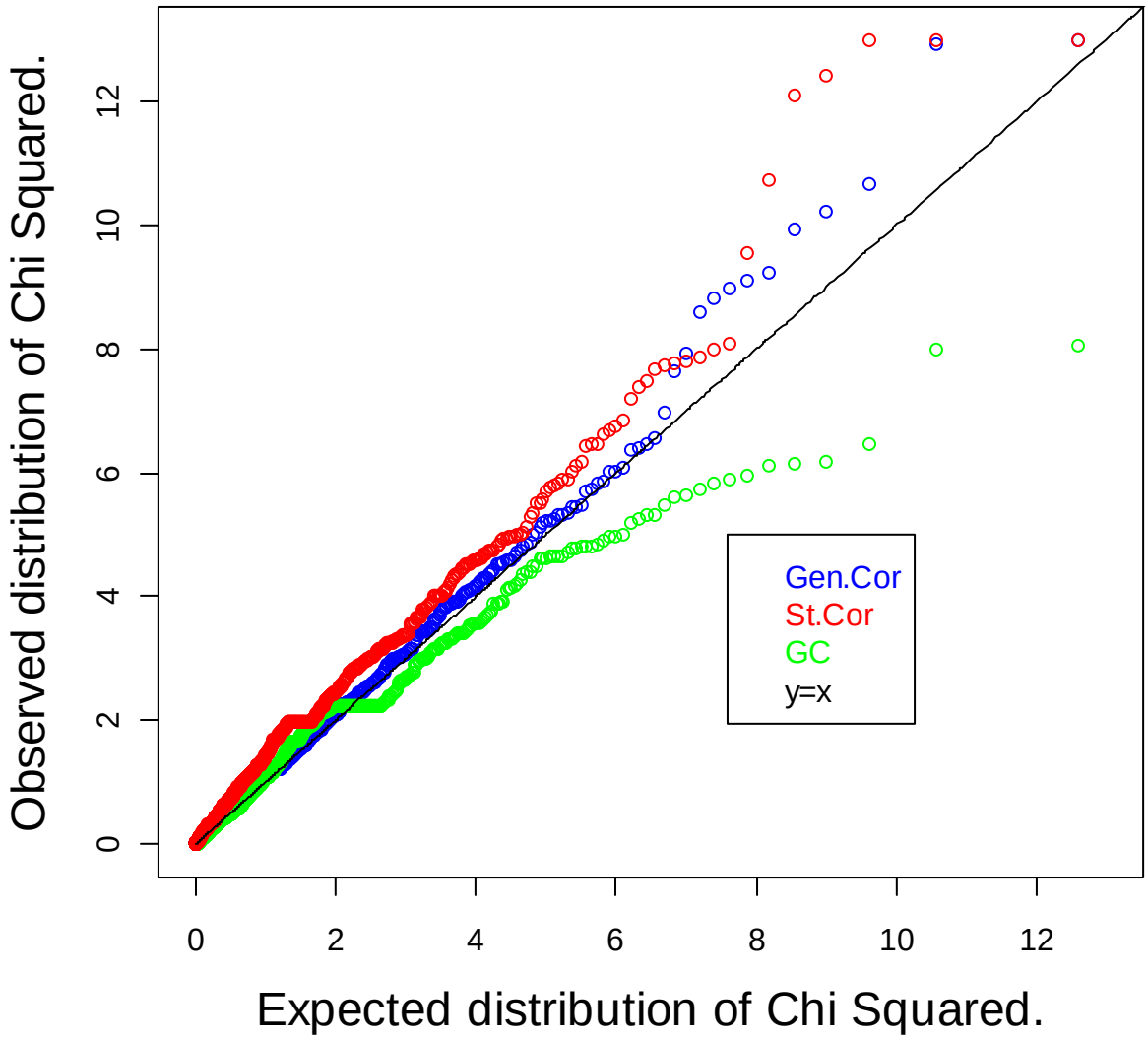


Figure 7-43 .Protein Content. QQ plot of Chi-Sq (1 d.f.) tests for association between marker and phenotype. Kinship has been modelled (MM) as statistical correlation (St.Cor) or genetic correlation (Gen.Cor) based on standardised genotype. MAF and maximum between marker R2 has been adjusted to the optimum values given in Table 7-13. Genomic control following an association test without adjustment for kinship is presented for comparison. Values greater than 13 have been set to 13.

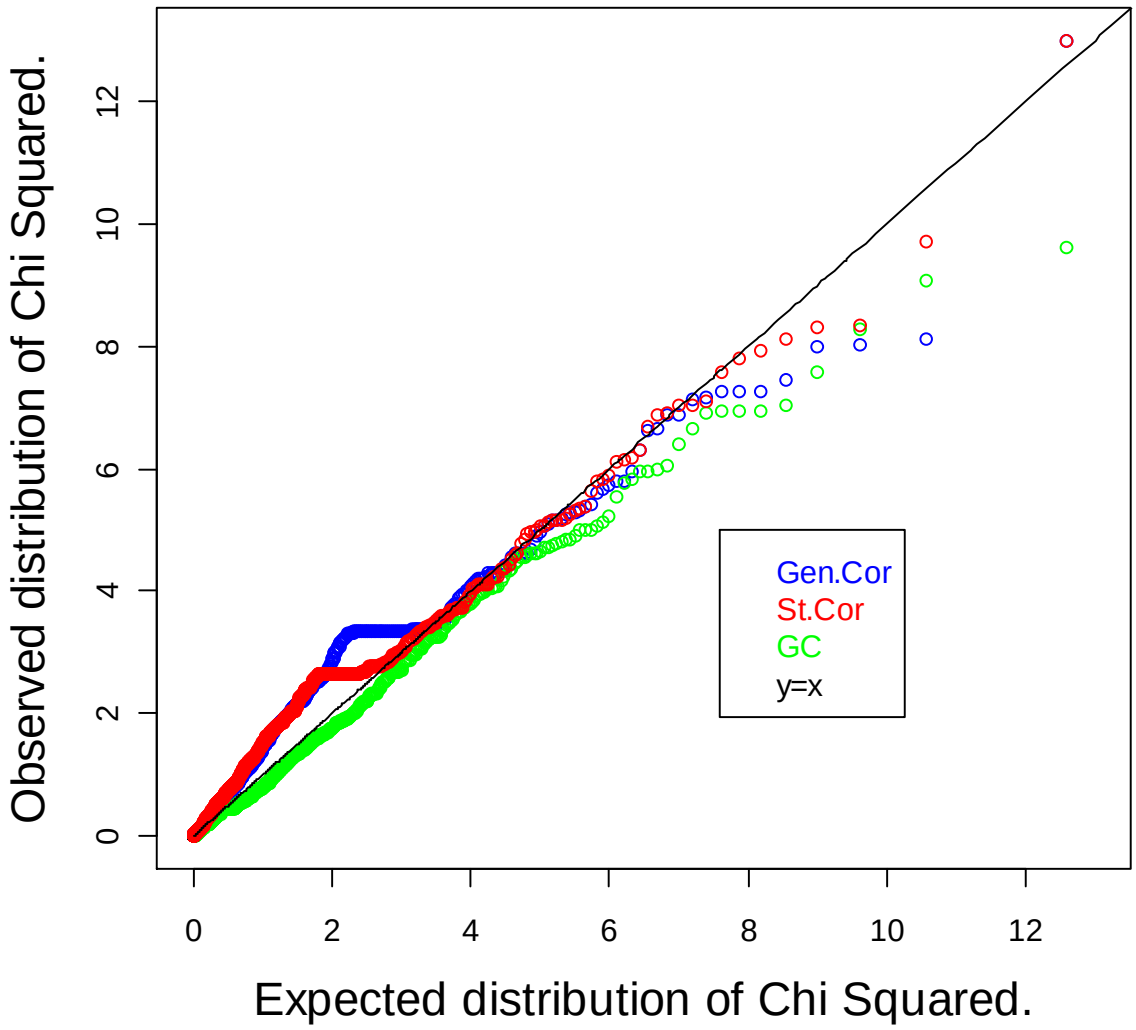


Figure 7-44 . Specific Weight. QQ plot of Chi-Sq (1 d.f.) tests for association between marker and phenotype. Kinship has been modelled (MM) as statistical correlation (St.Cor) or genetic correlation (Gen.Cor) based on standardised genotype. MAF and maximum between marker R2 has been adjusted to the optimum values given in Table 7-13. Genomic control following an association test without adjustment for kinship is presented for comparison. Values greater than 13 have been set to 13.

Considering each trait in turn. **Yield** (Figure 7-41) exhibits some general inflation of the test statistic when K_{stat} is used to measure kinship, this problem is less evident with the $K_{gen.}$ Both mixed models leave some evidence of association in that there are a group of points above the y=x line at high chi-square. GC, in contrast, shows no evidence of association for any marker. This is consistent with power simulations. In the case of **HFN** (Figure 7-42) both MM analyses offer good control of inflation and leave a group of potentially associated markers; some signal of association can be seen with GC but the signal is weaker than that from the

MM. Analysis of the **protein content** trait (Figure 7-43) shows control of inflation by both MM methods and a group of potentially associated markers is evident using K_{stat}. However, neither the K_{gen} nor GC models show any excess of significance. Finally, **specific weight** (Figure 7-44) exhibits no evidence of association with any of the three methods.

The inflation of the test statistic following mixed model analysis for each trait is summarised in Table 7-14.

Table 7-14. Inflation factor (median of observed chi-squared (1d.f.) divided by 0.455) for wheat phenotypes following association analysis using the MM.

Trait	Association without incorporation of a kinship model	MM K_{stat}	MM K_{gen}
Yield	7.67	1.45	1.07
HFN	2.50	1.24	1.00
Protein	6.80	1.28	1.11
Specific Weight	1.68	1.22	1.22

7.5.1 Significance of association.

As previously the significance of association has been estimated in terms of q-value. Plots of -log q-value over markers are presented in Figure **7-45** to Figure **7-48**.

It is evident (Figure 7-44 and Figure **7-48**) that there are no associations with **specific weight**. Considering the other traits however, there are modest peaks which may represent useful discoveries. In the case of **yield** two kinship models give different outcomes, one peak is shared by both but the $\mathbf{K}_{stat}$ model finds additional peaks across the genome, these may to some extent reflect the residual inflation seen in Figure 7-41 and Table 7-14 but it is possible they are genuine associations. I note them for further consideration. In addition to the mapped markers depicted in Figure **7-45** there are a number of markers which are unmapped or only assigned to a linkage group which are also significantly associated with yield. I summarise the 20 most significant associations together with cross reference to previous methods in Table 7-15.

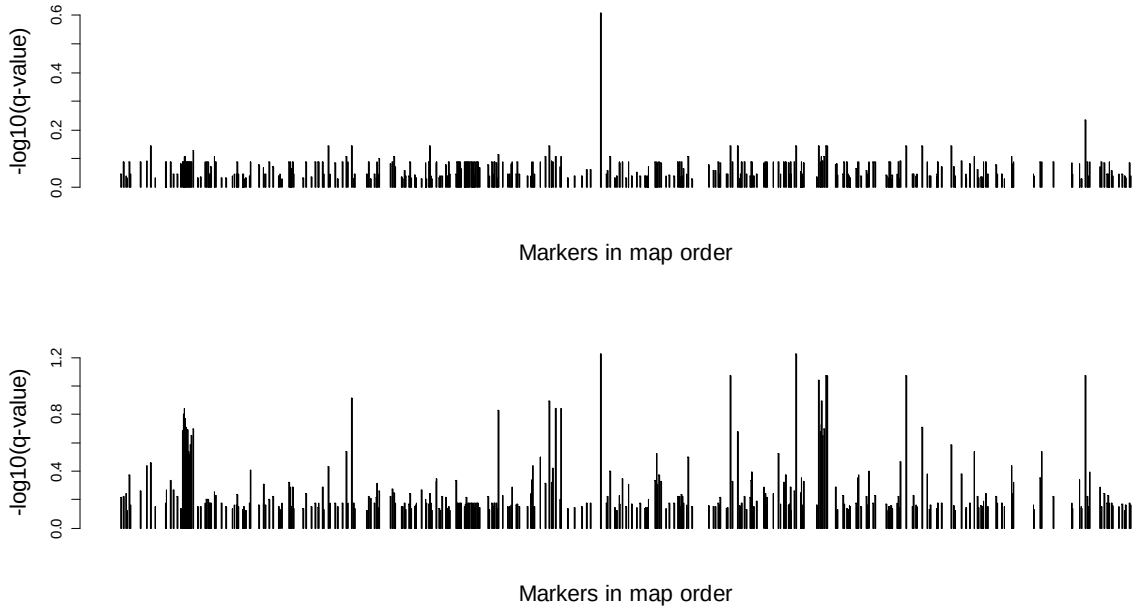


Figure 7-45. Yield. Association with mapped markers. Upper graph shows results from a MM using K_{gen}, lower graph MM using K_{stat}. K matrices determined using all markers skimmed to optimum values given in Table 7-13.

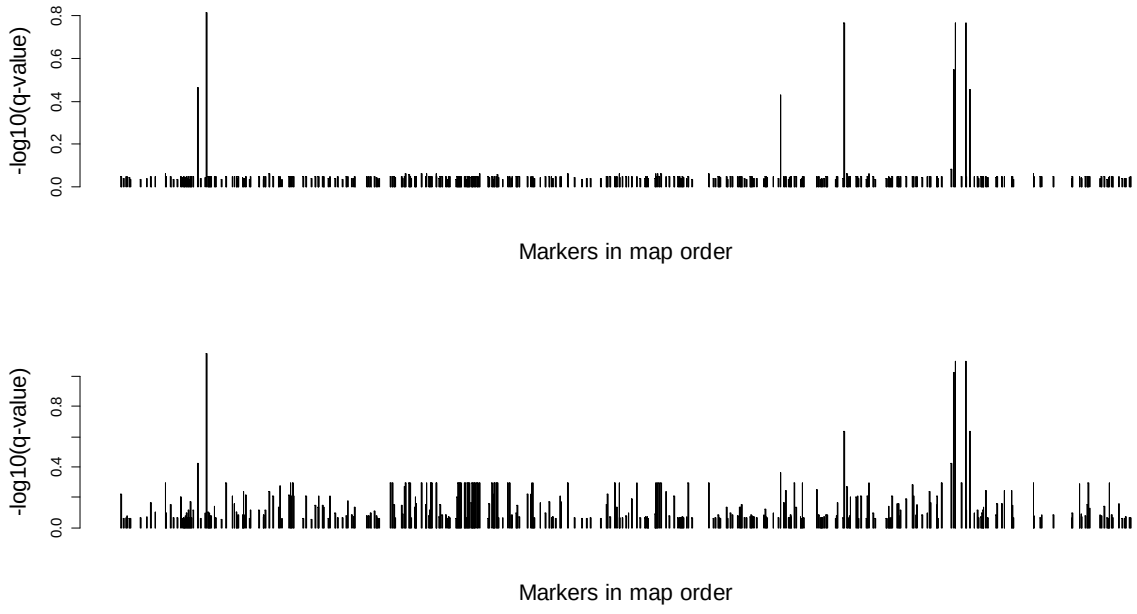


Figure 7-46. HFN. Association with mapped markers. Upper graph shows results from a MM using K_{gen}, lower graph MM using K_{stat}. K matrices determined using all markers skimmed to optimum values given in Table 7-13.

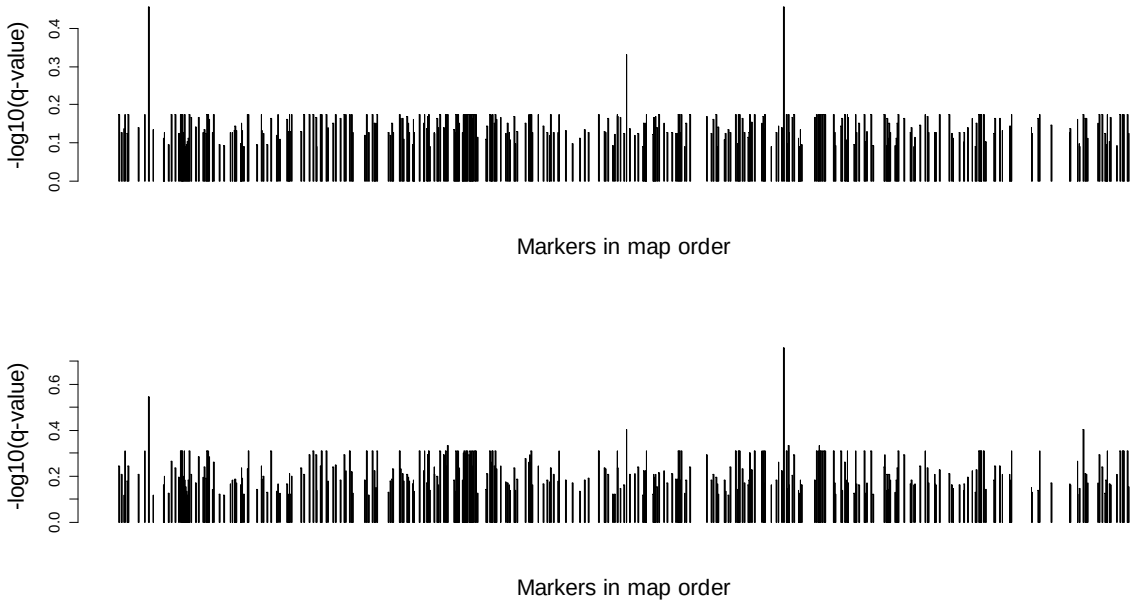


Figure 7-47. Protein content. Association with mapped markers. Upper graph shows results from a MM using K_{gen}, lower graph MM using K_{stat}. K matrices determined using all markers skimmed to optimum values given in Table 7-13.

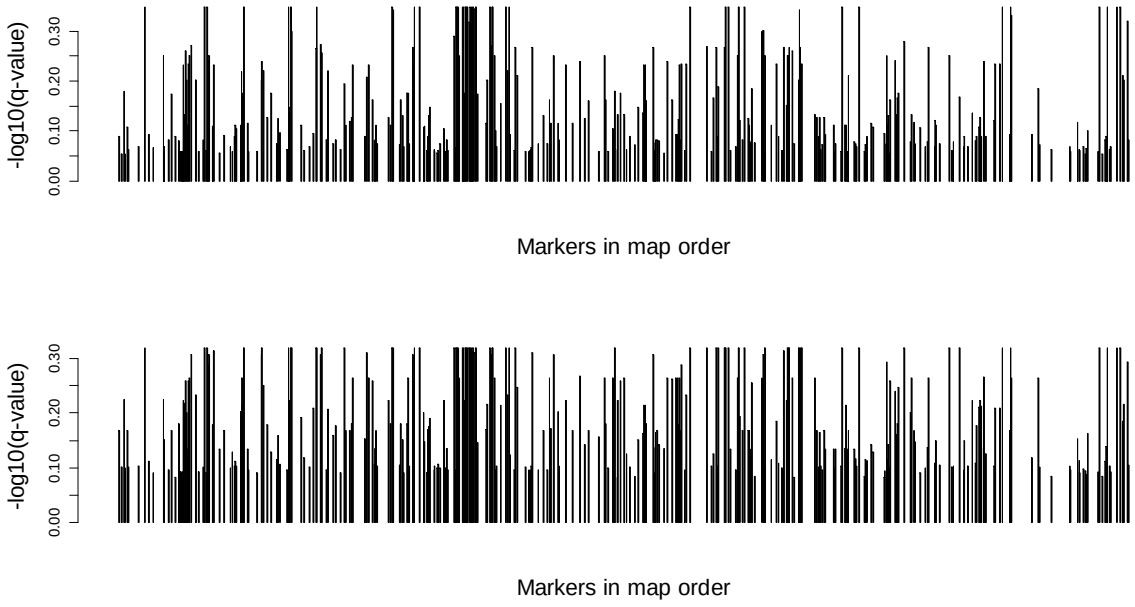


Figure 7-48. Specific weight. Association with mapped markers. Upper graph shows results from a MM using K_{gen}, lower graph MM using K_{stat}. K matrices determined using all markers skimmed to optimum values given in Table 7-13.

Table 7-15. 20 markers most significantly associated with yield (MM analysis K_{stat}). (Key to published QTL references in Table 7-19).

Marker	cM	C'some	Effect size	h^2	P_{var} Acc	q-value	SA(6)	Naïve	PCA(12)	MM K_{gen}	Publ' QTL
wPt-2397	105.8	2B	-0.61	0.91	0.09	0.04	-	-	-	+	8,9
wPt-2989	Unmapped	Unassigned	-0.59	0.91	0.09	0.04	-	+	-	+	
wPt-0948	Unmapped	Unassigned	-0.60	0.91	0.09	0.04	-	+	-	+	
wPt-3754	Unmapped	Unassigned	-0.61	0.91	0.09	0.04	-	+	-	+	
wPt-6108	Unmapped	Unassigned	-0.59	0.91	0.09	0.04	-	+	-	+	
wPt-0391	68.7	4B	-0.90	0.91	0.07	0.08	-	-	-	+	
wPt-7323	25.8	5D	-0.47	0.91	0.07	0.08	-	-	-	-	18
wPt-1922	12.9	6B	-0.70	0.91	0.07	0.08	+	-	-	-	
wPt-3130	12.9	6B	-0.71	0.91	0.07	0.08	+	-	-	-	
wPt-9015	Unmapped	Unassigned	-0.70	0.91	0.07	0.08	+	-	-	-	
wPt-4386	Unmapped	Unassigned	-0.70	0.91	0.07	0.08	+	+	-	-	
wPt-6563	Unmapped	Unassigned	-0.70	0.91	0.07	0.08	+	-	-	-	
wPt-4720	12.5	6B	-0.69	0.91	0.07	0.09	+	-	-	-	
wPt-2448	Unmapped	Unassigned	0.45	0.90	0.08	0.09	-	-	-	-	
wPt-1238	Unmapped	Unassigned	-0.66	0.94	0.02	0.10	-	-	-	-	
wPt-5383	Unmapped	Unassigned	-0.68	0.91	0.07	0.10	+	-	-	-	
wPt-2696	Unmapped	Unassigned	-0.43	0.93	0.03	0.12	-	-	-	+	
wPt-1553	47.4	7B	-0.73	0.90	0.08	0.12	-	-	-	-	
wPt-7150	12.7	6B	-0.59	0.90	0.07	0.14	+	-	-	-	
wPt-3565	Unmapped	Unassigned	0.46	0.91	0.07	0.14	-	-	-	-	

Table 7-16. Mapped markers associated with yield following MM analysis q-value≤0.3. (Key to published QTL references in Table 7-19).

Marker	cM	C'some	Effect size	h^2	P_{var} Acc	q-value	SA(6)	Naïve	PCA(12)	MM K_{gen}	Publ' QTL	Hit	Miss
wPt-3753	13.42	1B	0.46	0.91	0.04	0.16	-	-	-	-	-		1
wPt-8328	76.43	2A	0.36	0.91	0.05	0.21	+	-	-	-	-		1
wPt-3611	76.53	2A	0.39	0.90	0.06	0.17	+	-	-	-	-		
wPt-5027	76.65	2A	0.41	0.91	0.06	0.16	+	-	-	-	-		
wPt-1368	76.70	2A	0.45	0.91	0.05	0.17	+	-	-	-	-		
wPt-6148	76.97	2A	0.36	0.91	0.05	0.20	+	-	-	-	-		
wPt-9624	77.82	2A	0.38	0.90	0.05	0.23	+	-	-	-	-		
wPt-6207	78.09	2A	0.38	0.91	0.04	0.28	+	-	-	-	-		
wPt-7626	78.50	2A	0.38	0.90	0.06	0.26	+	-	-	-	-		
wPt-2087	79.51	2A	0.34	0.91	0.05	0.18	-	-	-	-	-		
wPt-8404	30.27	2B	-0.37	0.93	0.02	0.16	-	-	-	-	8,9	1	
wPt-9402	49.10	2B	-0.46	0.91	0.06	0.16	-	-	-	-	8,10	1	
wPt-3983	50.58	2B	-0.51	0.93	0.02	0.16	-	-	-	-	8,10		
wPt-2397	105.80	2B	-0.61	0.91	0.09	0.04	-	-	-	-	8,9	1	
wPt-2397	105.80	2B	-0.52	0.91	0.09	0.25	-	-	-	+	8,9		
wPt-2119	29.37	3B	0.45	0.94	-0.01	0.22	-	-	-	-	-		1
wPt-2689	89.55	3D	-0.51	0.91	0.04	0.28	+	-	-	-	15	1	
wPt-0391	68.72	4B	-0.90	0.91	0.07	0.08	-	-	-	-	-		1
wPt-8292	90.15	4B	-0.46	0.92	0.04	0.16	-	-	-	-	18	1	
wPt-1250	43.68	5B	-0.53	0.92	0.02	0.26	-	-	-	-	8	1	
wPt-7323	25.79	5D	-0.47	0.91	0.07	0.08	-	-	-	-	18	1	
wPt-4255	30.25	6A	0.41	0.89	0.08	0.16	+	-	-	-	-		1
wPt-4720	12.51	6B	-0.69	0.91	0.07	0.09	+	-	-	-	-		1
wPt-7777	12.61	6B	-0.56	0.91	0.04	0.20	-	-	-	-	-		
wPt-6994	12.67	6B	-0.63	0.91	0.05	0.16	-	-	-	-	-		
wPt-7150	12.74	6B	-0.59	0.90	0.07	0.14	+	-	-	-	-		
wPt-3304	12.77	6B	-0.59	0.91	0.04	0.18	-	-	-	-	-		
wPt-4867	12.81	6B	-0.59	0.91	0.05	0.16	-	-	-	-	-		
wPt-1922	12.85	6B	-0.70	0.91	0.07	0.08	+	-	-	-	-		
wPt-3130	12.85	6B	-0.71	0.91	0.07	0.08	+	-	-	-	-		
wPt-1553	47.45	7B	-0.73	0.90	0.08	0.12	-	-	-	-	-		1
wPt-7887	73.11	7B	-0.47	0.91	0.05	0.16	-	-	-	-	18	1	
wPt-0884	84.28	7B	-0.37	0.92	0.03	0.28	-	-	-	-	18		

Some markers were found to be co-linear with published QTL for the trait, where this is the case I give the code to the literature references used in the full summary given later (foot of Table **7-19**). Table 7-17 gives the same information but for markers which showed significant association (p-value < 0.3) and for which a genetic map position was available. In addition I show a count of 'Hits' where a marker, or group of markers, are collinear with a published SNP and 'Misses' where markers or groups of markers were found not to be co-linear with published QTL.

Figure **7-46** shows that for **HFN** the patterns of association found using the MM with both K_{stat} and K_{gen} are similar. Only 14 markers are significant at Q≤0.3 and these are summarised in Table **7-17** together with cross reference to previous methods and a summary of coincidence with published QTL. References for published QTL are given below Table 7-19.

Table 7-17. All markers associated with Hagberg falling number, MM analysis q-value≤0.3. (Key to published QTL references in Table 7-19).

Marker	cM	C'some	Effect size	QTL h^2	P_{var} Acc	q-value	SA	Naïve	MM K_{gen}	Publ' QTL	Hit	Miss
wPt-0126	Un-mapped	Un-assigned	-41.38	0.45	0.14	0.07	+	+	+			
wPt-3785	Un-mapped	Un-assigned	-40.25	0.45	0.13	0.08	+	+	+			
wPt-4319	Un-mapped	Un-assigned	-38.01	0.47	0.11	0.09	+	+	+			
wPt-8596	Un-mapped	Un-assigned	-50.46	0.59	0.08	0.17	-	-				
wPt-9013	Un-mapped	Un-assigned	-34.01	0.48	0.10	0.23	-	+				
wPt-4644	Un-mapped	Un-assigned	-35.07	0.46	0.11	0.23	+	+				
wPt-4319	Un-mapped	Un-assigned	-37.55	0.51	0.12	0.28	+	+				
wPt-1480	157.1	2A	83.08	0.64	0.04	0.07	-	-	+	21	1	
wPt-1241	35.8	6B	-25.91	0.52	0.07	0.23	+	-	+	20	1	
wPt-6320	84.6	7B	-39.48	0.46	0.13	0.08	+	+	+	1,12	1	
wPt-9813	84.6	7B	-38.81	0.46	0.12	0.09	+	+	+	1,12		
wPt-9813	84.6	7B	-38.34	0.50	0.13	0.28	+	+		1,12		
wPt-1533	85.2	7B	-39.85	0.46	0.12	0.08	+	+	+	1,12		
wPt-5069	85.6	7B	-36.39	0.45	0.11	0.23	-	+		1,12		

Consistent with the P-P plots, the signal for association with **protein** is weak for both kinship models (Figure **7-47**). However, two mapped and five unmapped markers were significant at Q ≤ 0.3 and, as previously, these are summarised in Table 7-18.

Table 7-18. All markers significantly associated with Protein, MM analysis q-value≤0.3. (Key to published QTL references in Table 7-19).

Marker	cM	C'some	Effect size	QTL h^2	P_{var} Acc	q-value	SA	Naïve	EIG	MM K_{gen}	Publ' QTL	Hit	Miss
wPt-7340	Un- mapped	Unassigned	0.52	0.71	0.13	0.05	-	-	-	-	-		
wPt-8355	Un- mapped	Unassigned	0.52	0.71	0.13	0.05	-	-	-	-	-		
tPt-2230	Un- mapped	Unassigned	0.75	0.76	0.10	0.05	-	-	-	+	-		
wPt-6191	Un- mapped	Unassigned	0.39	0.74	0.10	0.10	+	-	-	-	-		
tPt-4614	Un- mapped	Unassigned	0.67	0.79	0.05	0.10	-	-	-	+	-		
wPt-9757	101.30	1A	0.33	0.75	0.07	0.28	-	-	-	-	8	1	
wPt-2707	65.06	5B	0.37	0.74	0.09	0.18	-	-	-	-	8	1	

7.6 Significance of discoveries

Demonstrating that markers are significantly associated with a trait is a useful outcome if one is searching for QTL which may ultimately have practical application. However, significant association only allows the null hypothesis to be rejected; it provides no information as to the correct alternate hypothesis to accept.

In order to estimate how likely it was that the reported discoveries resulted from a genetic effect on the trait I considered the QTL already published for the traits tested. These QTL had come from biparental mapping studies and were thus immune to the confounding effects of population structure and kinship. If the association discoveries were coincident with the discoveries from bi-parental mapping it would be consistent with the alternate hypothesis that the discoveries by association resulted from an LD signature between these QTL and the significant markers.

From a literature review () I found QTL for HFN, Protein and Yield. I mapped these using the locations given by the authors and used my own consensus map to overlay the positions of the markers showing significant association. I noted many occasions when more than one author had discovered a QTL mapping to a particular location, I treated these as a single QTL. summarises the numbers of QTL and the discoveries by association.

Table 7-19 References to QTL identified for yield, protein and HFN identified in a literature search. (Key to published QTL references below table).

Chromosome	Yield	Protein	HFN
1A	9, 18	(4, 8)	21
1B	9, 18		12, 21
1D			21
2A	(9, 14, 15)	(8, 16), 16	21, 21
2B	(15, 18), (8, 9)	16	21
2D	8, (8,14)	16	(11, 12)
3A	3, 3	(8, 12)	7, 21
3B	(8, 9), (9, 14)	20	21, (7, 7 , 11, 21)
3D	15, 15, 18	16, 16	(8, 11)
4A	16, 18, (8, 18)	12, (8, 12, 16, 16)	1, 21
4B	18, (2, 8, 18)	(2, 4)	12, 21
4D	9, (9, 15, 18)	(4, 8)	
5A	(2, 10), (10, 18)	2, (4, 20)	7
5B	8, 9, 18,	4, 8	21
5D	18, 18	4, 12	
6A	(2, 19), 14	2, 8	
6B	2, 18	(2, 5, 16), 16	21
6D			
7A	2, 17	(2, 8, 16), 4	
7B	18, 18	(2, 4, 12)	(12, 21), (1, 12), 21, 21
7D	8, 8	4, 12, 16	

Note: References bracketed together identify co-linear QTL. Multiple identical references in one chromosome indicate the authors identify more than one QTL.

Key to References:

1.Batey *et al.* (2001); 2.Blanco *et al.* (2002); 3.Campbell *et al.* (2003); 4.Campbell *et al*. (2001); 5.Distelfield *et al.* (2008); 6.Distelfield *et al.* (2006); 7.Groos *et al.* (2002); 8.Groos *et al*. (2003); 9.Huang *et al.* (2004); 10.Kato *et al.* (2000); 11.Kulwal *et al.* (2004); 12.Kunert *et al.* (2007); 13.Li *et al.* (2004); 14.Li *et al.* (2007); 15.McCartney *et al.*(2005); 16.Prasad *et al.* (2003); 17.Quarrie *et al.*(2006); 18.Quarrie *et al.* (2005); 19.Snape *et al.* (2007); 20.Sun *et al.* (2008); 21.Zanetti *et al.* (2000).

Table 7-20. Numbers of published QTL for the studied traits compared with numbers of discoveries by association.

Trait	Published QTL (duplicate discoveries counted singly)	Total discoveries by Association (closely adjacent discoveries counted singly)	Association coincident with a QTL.	Association not coincident with any QTL.
HFN	24	3	3	0
Protein	29	2	2	0
Yield	40	15	8	7

I tested the null hypothesis that the locations of the discoveries by association were independent of the locations of the published QTL by randomly placing the appropriate number of association discoveries on the genetic map and counting occasions when they were coincident with published QTL. This process was repeated as a Monte Carlo simulation (10,000 cycles) in MS Excel using the PopTools add in (from: http://www.cse.csiro.au/poptools/index.htm). The simulations were repeated over a range of average QTL widths, results are summarised in .

Table 7-21 Statistical significance[1] of the observed number of discoveries by association which are coincident with published QTL. (Estimated from 10,000 Monte Carlo Simulations).

Width of QTL (cM)	HFN	Protein	Yield
5	***	***	***
9	***	**	***
13	***	**	**
17	**	•	•
21	**	•	NS
25	**	NS	NS
29	**	NS	NS
33	•	NS	NS
37	•	NS	NS
41	•	NS	NS
45	NS	NS	NS

[1] NS = not significant, *** =P≤0.001, ** =P≤0.01, *=P≤0.05, •=P≤0.1.

7.7 Discussion.

The statistical analysis allows the allocation of results between alternate and null hypotheses. The final role for the researcher is to attempt to define the correct alternate hypothesis or hypotheses.

In this experiment power had previously been determined to be relatively low for markers accounting for 5-10% of the phenotypic variation (which seems a typical value for the more significant markers; Table 7-15 to Table 7-18). Specifically, when detecting the location of the causative marker to within 10cM, power is ~40% for yield and HFN, ~45% for Protein and ~20% for specific weight. Thus, depending on trait, 55 – 80% of real associations will have been attributed to the null when the significance threshold was $q \leq 0.3$. By definition (with this threshold), 30% of reported discoveries must come from the null hypothesis. Nonetheless if un mapped associations are considered there remain a number of discoveries which can be allocated to an alternate hypothesis.

At least two alternate hypotheses seem possible (a) Statistically significant association is due to linkage. (b) Statistically significant association is due to uncontrolled latent variables in the experiment. To some extent it is possible to quantify the risk from (b) since in the power simulations with the simulated phenotype the empirical false discovery rate may be of the order of 50% (Figure 6-18), these false discoveries comprise markers falling under (b) since the trait was fully simulated.

Given these approximate values, may be re-drawn as Table 7-22. Since power is low there are many discoveries which have been wrongly allocated to the null hypothesis, unfortunately these cannot be recovered without improving the experiment or further relaxing the significance threshold. However, empirically 45% of all discoveries are expected to be false and some discoveries can be allocated to hypothesis (a) since they are coincident with published QTLs. It is then possible to calculate the expected number of false discoveries among these unallocated discoveries. illustrates that the experimental outcome has left no mapped associations unaccounted for by the estimated statistical power. It should be noted, however, that this calculation relates to mapped markers, it is also true of the unmapped set. If these were mapped then approximately half of the loci they represent (note that some markers may map to the same locus and so would count only once) would, from simulation, be expected to be genuine associations.

Association with specific weight proved impossible to detect in this experiment despite the trait heritability being similar to that observed for protein. However power estimated from the perturbed phenotype was only 20-30% compared to over 40% for protein suggesting association with the trait would be hard to detect. One explanation in the association study might be that the trait lacks QTL of large effect; however this would not have resulted in low power in the perturbed phenotype. If the analytical data were inaccurate this would reduce power but the estimate of h^2 for specific weight made from analyses of the trials data is 90% (similar to the other traits) which suggests there has been reasonable control of experimental error. It is possible that specific weight is varying in such close concert with a structural feature of the genotypes that control of structure in the model removes association signals as well; although possible, this is not really plausible in this case since the effect is not evident in Figure 7-44. It is tempting to dismiss the failure as inevitable in an underpowered experiment and this may well account for the failure to detect association but the unanswered question is to why the power to detect a perturbed phenotype is also low?

Table 7-22 Allocating discoveries to alternate hypotheses and attempting to quantify unallocated true discoveries.

Trait	Published QTL (duplicate discoveries counted singly)	Total discoveries by Association (closely adjacent discoveries counted singly)	1-power	QTL wrongly attributed to the null	Potentially true discoveries discarded due to low power	Association coincident with a published QTL. (Discoveries which are likely to be due to genetic proximity).	Association not coincident with any QTL. (Discoveries for which the correct alternate hypothesis is yet to be determined).	False discoveries expected since q<=0.3	False discoveries expected since empirical FDR is 0.45	Excess of false discoveries since empirically we observe 45% false discovery in simulation.
HFN	24	3	0.6	5	3	3	0	1	1	0
Protein	29	2	0.55	4	2	2	0	1	1	-1
Yield	40	15	0.6	25	14	8	7	5	7	0

8 Association Mapping: Barley Panel.

[Note that the positions of associations for yield described here are proprietary to the AGOUEB consortium and permission has not been given for their release. For this reason the chromosome identities have been re-coded (A-G) in random, but consistent, order. It is likely that the information will be cleared for release in future in which case detail could be supplied (jon.white@niab.com)].

The analyses presented here contributed to the recent publication (Cockram *et al.* 2010) in which multiple trait/marker associations are reported and one trait is further analysed to uncover a potential causative polymorphism. In this publication I supplied power calculations, association mapping analysis, confirmation that the assumption of normality of residuals in the MM could be set aside in order to calculate marker to marker association and the estimate of LD over genetic map.

8.1 Marker and variety panels

Marker and variety panels used in the initial association analysis are as described in Chapter 5. For one trait additional markers were developed in a region containing a candidate gene, these are described in 8.9.

8.2 Statistical Power.

Initial analysis of the association panel in Chapter 6 suggests the statistical power to detect highly heritable Mendelian traits was in the range 80-95%. Quantitative traits, however, were likely to be harder to detect; if a QTL explained 10-15% of phenotypic variation power would be ~ 25-50%.

8.3 Phenotypes

The phenotypes available for the barley panel comprised yield together with thirty-four botanical characters used for variety classification. Some of the phenotypes are simple binary traits, some are discontinuous ordinal trait and others have a quantitative character. Table 8-1 summarises the traits. Detailed descriptions of the majority of the traits can be found at http://www.cpvo.europa.eu/documents/TP/agricoles/TP_019-2rev_HORDEUM_VULGARE.pdf where the CPVO reference number given in Table 8-1 identifies the description. Distribution of phenotype scores in the panel are summarised as histograms in Appendix 6.

Of particular note are phenotypes relating to anthocyanin colouration in diverse plant tissues; this natural pigment is produced in response to stress. However, some varieties never produce anthocyanin (anthocyanin incompetent) and others express the trait in some tissues but not others. Additionally the level of anthocyanin production is genetically determined such that among competent varieties some produce the pigment in abundance whilst others do so only weakly.

Table 8-1a. Summary of Binary Phenotypes.

Phenotype name	CPVO No.[1]	Phenotype name	CPVO No.
Anthocyanin competence	Derived[2]	Spiculation of awn margins	N/A
Anthocyanin colouration of auricles	N/A	Hairs in the ventral furrow	25
Anthocyanin colouration of awn tips	N/A	Adherence of grain husk	22
Hairiness of the lower leaf sheaths	2	Disposition of lodicules	26
Row number (6/2)	11	Seasonal type (W/Sp)	28
Rachilla hair type	21		

Table 8-1b. Summary of Discontinuous Ordered Phenotypes.

Phenotype name	CPVO No.	Phenotype name	CPVO No.
Intensity of anthocyanin colouration of auricles	3	Development of the sterile spikelet	18
Intensity of anthocyanin colouration of awn tips	7	Length of glume c.f. grain	20
Anthocyanin colouration of grain lemma nerves	23	Ear length. Awn length	20
Plant growth habit	1	Plant length	10
Frequency of plants with re-curved leaves	4	Collar type	N/A
Time of ear emergence	6	Ear density	13
Ear glaucosity	8	Ear shape	12
Glaucosity of the flag leaf sheath	5	Length of first rachis segment	16
Ear attitude	9	Curvature of first rachis segment	17
Sterile spikelet attitude	19	Colour of the aleurone layer	27
Shape of the tip of the sterile spikelet	N/A	Spiculation of inner lateral nerves	24

Table 8-1c Summary of Continuous (quantitative) Phenotypes.

Phenotype name
Yield (t/ha)

[1] CPVO reference number of the trait. N/A are traits not specified by SPVO but which are collected by NIAB.

[2] Derived phenotype; varieties which produced anthocyanin in any tissue type were scored ‘1’ varieties which never produced anthocyanin were scored ‘0’.

The yield phenotype was estimated using the mixed effects model (See; **Appendix 1** for outline explanation) in plot trials over multiple sites and seasons, the analyses were conducted by Jordi Comadran at the Scottish Crop Research Institute (SCRI, Dundee.) using the model:

Yield ~ Fixed effects: Y + T + R + Y.T.R Random effects: V + e

Where:

Y= Year, T= Trial site, R= Geographic region, V= Variety.

The phenotype data for the botanical characters were provided as variety descriptions without replicated data to support them. This does not mean the phenotypes resulted from a single observation but rather that the observations over multiple plants and seasons had been summarised as a single value used to describe the variety. This means that h^2 cannot be estimated from trials data for the botanical characters.

8.4 Association Mapping.

8.4.1 Effect of structure.

As has been discussed (5.1.4), structure within the barley panel relates primarily to the division between Spring and Winter types and within those to the division between 6-row and 2-row barley. These divisions correspond to separate breeding pools. Again as previously discussed, structural divisions may result in genome wide inflation of estimates of association of markers with trait whenever the average value of the trait and the allele frequencies of markers co-vary between these divisions.

A simple metric for inflation is the median of a 1d.f. association statistic, for example chi squared which, under the null hypothesis, has the expected value 0.455 and so an inflation factor is simply a ratio of the median observed in the experiment to this expectation (c.f. Genomic Control 3.3.5.1). A summary of inflation factor by phenotype is presented in Table 8-2; in this case chi-sq 1df has been approximated as a likelihood ratio test based on the change in deviance following the addition of the tested marker to a simple regression model.

It is interesting that some phenotypes show only modest inflation whereas others show a much larger deviation from the null. Some evidence of the source of this inflation can be found by comparing inflation factors for spring barley and winter barley yields which are 2.7 and 5.3 respectively but in the combined data set the phenotype has an inflation factor of 104.

Table 8-2. The inflation of the test statistic for association (chi-squared, 1df) for each phenotype.

Phenotype name	Inflation factor (1 = no inflation).	Phenotype name	Inflation factor (1 = no inflation).
Adherence of grain husk	1.04	Winter barley yield	5.31
Shape of the tip of the sterile spikelet	1.19	Intensity of anthocyanin colouration of auricles	6.09
Time of ear emergence	1.36	Awn length	7.94
Spiculation of awn margins	1.43	Anthocyanin colouration of grain lemma nerves	13.31
Frequency of plants with re-curved leaves	1.61	Colour of the aleurone layer	14.84
Ear shape	1.62	Ear glaucosity	16.81
Intensity of anthocyanin colouration of awn tips	1.65	Row number (6/2)	19.89
Length of glume c.f. grain	1.83	Development of the sterile spikelet	20.50
Length of first rachis segment	2.18	Sterile spikelet attitude	20.68
Ear length.	2.30	Hairs in the ventral furrow	21.83
Curvature of first rachis segment	2.37	Anthocyanin colouration of auricles	27.44
Plant length	2.66	Anthocyanin colouration of awn tips	27.77
Spring barley yield	2.71	Spiculation of inner lateral nerves	29.69
Collar type	2.89	Anthocyanin competence	30.09
Ear density	3.03	Plant growth habit	42.52
Disposition of lodicules	3.09	Hairiness of the lower leaf sheaths	80.93
Ear attitude	3.13	Seasonal type (W/Sp)	95.69
Glaucosity of the flag leaf sheath	3.73	Winter and winter barley yield	104.34
Rachilla hair type	4.64		

8.4.2 Modelling structure.

As already discussed (3.3.3), there are two views of population structure: (A) a sub-population structure view in which partial or complete isolation of ancestral groups has allowed stochasitic (and other) processes to change allele frequencies at many loci within these groups. (B) a kinship view in which genetic similarity between individuals can be explained in terms of pedigree. There is a less well reported third perspective, to which I subscribe, in which these two models are viewed as extremes of a continuum.

Over these multiple phenotypes it is possible crudely to test these two models (A and B) as an explanation for genome wide inflation.

Structure is commonly modelled as the first (~20) principal coordinates of a correlation or covariance matrix, or in terms of **Q** from STRUCTURE. Kinship can be variously estimated; here I use pair-wise correlation between varieties calculated from standardised genotype. In all cases structure and kinship were estimated in a sub-set of markers selected to be spaced at ~2cM intervals.

In the case of structure, each variety has a position on each of the principal component axes or a fractional membership of each of *k* sub-populations; these may be used as co-variates in a model[1] to predict phenotype. Correlation (r) estimated from this prediction is a standardised measure of the co-variation of structure and phenotype. In the case of kinship the situation is slightly different and an estimate of similarity between variety pairs (r) can be correlated with the corresponding phenotypic distance; in this case there will be a negative correlation between the two measures. In all cases the explanatory power of the structure/kinship model should correlate with the inflation of the test statistic given in Table 8-2.

Results are presented in Figure 8-1 to Figure 8-3 and show that over the 37 traits the stronger the association of phenotype with the structure/kinship model the greater the observed inflation of the test statistic in a simple marker-trait association. This is entirely consistent with the generally supposed origins of spurious association in association mapping already discussed(3.2 and 3.3). Additionally, it is interesting to note that in Figure 8-1 and Figure 8-2 traits exhibiting generally low inflation are correlated with the structure model in the range r = 0.05 – 0.4 suggesting that the structure model may be an unreliable basis for deflating the test statistic when the level of inflation is modest. Figure 8-3 shows that kinship is an excellent predictor of pair-wise phenotypic distance for traits where inflation of the test statistic is high. Where inflation is low kinship gives consistently poor prediction of phenotypic distance (i.e. pairwise difference in phenotype) suggesting that kinship may be a more consistent basis for correction when the inflation factor is lower.

[1] Generalised linear model with a Gaussian link function for continuous traits and a logit link function for binary traits.

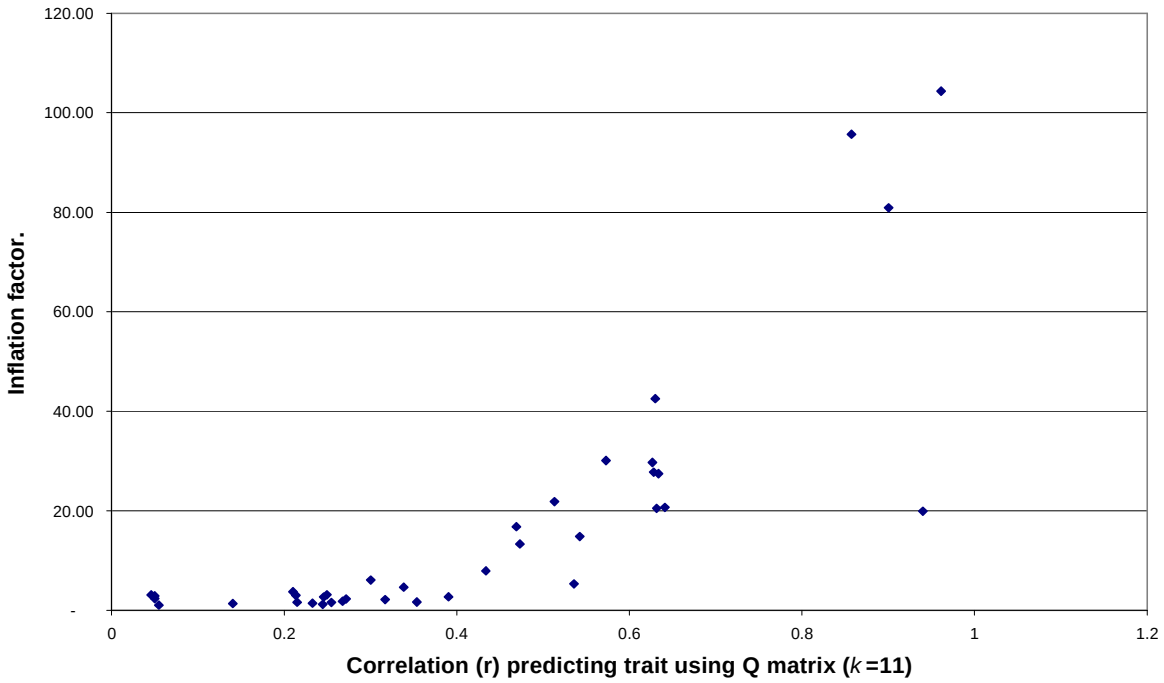


Figure 8-1. Correlation from a regression model in which each of 37 phenotypes are predicted using Q (k=11) against the inflation, relative to the expectation under the null, of the median of the test statistic (chi-sq, 1df) for association in a simple model associating marker with trait. The point at about (0.95,20) which appears out of trend is the 'number of rows' character.

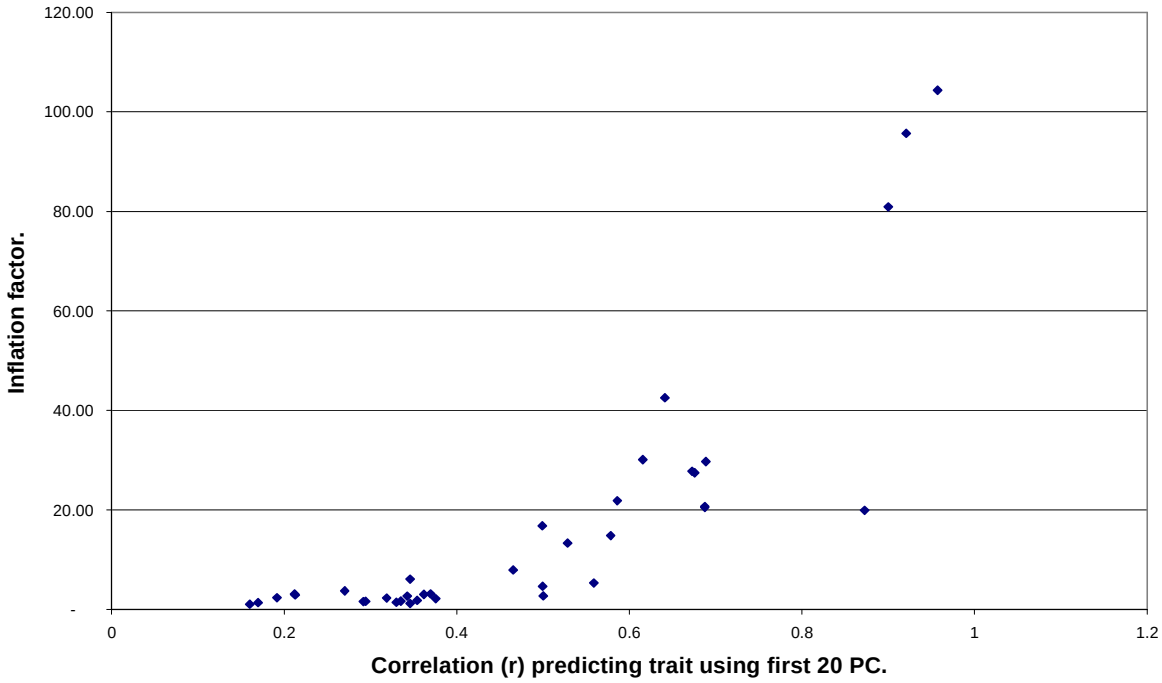


Figure 8-2. Correlation from a regression model in which each of 37 phenotypes are predicted using the first 20 principal coordinates of a correlation matrix against the inflation, relative to the expectation under the null, of the median of the test statistic (chi-sq, 1df) for association in a simple model associating marker with trait. The point at about (0.85,20) which appears out of trend is the 'number of rows' character.

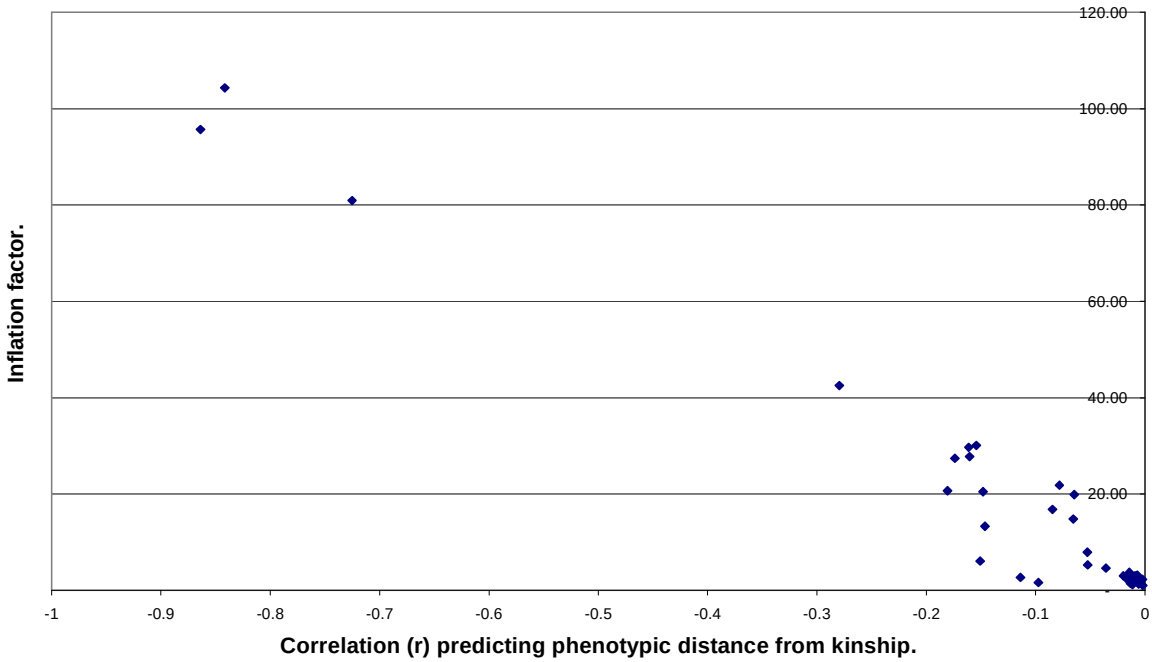


Figure 8-3. Correlation between phenotypic distance and kinship for each of 37 phenotypes against the inflation, relative to the expectation under the null, of the median of the test statistic (chi-sq, 1df) for association in a simple model associating marker with trait.

In contrast to the structure models which define the relationships between varieties in terms of dimensions of variation or inferred historic populations, the kinship model is a continuous scale of proximity in which strong but localised features in the genotype are not interpreted in terms of sub-population divisions but contribute to an average score of relatedness. In addition, hard genetic signatures (for example groups of correlated markers), possibly closely related to the causative locus, for many of the 'low inflation' traits may be interpreted as weak structural divisions but their signature is obscured in the kinship score. This contrasting view would account for the ability of structure models to predict some phenotypes even when the inflation factor was low (since features local to the QTL give a weak structural signature) and also account for the contrasting inability of kinship to do the same.

For reference, Table 8-3 lists the traits which appear likely to be only modestly affected by spurious association.

Table 8-3 Traits with low inflation of the test statistic and low correlation between kinship and phenotypic distance.

Phenotype	Correlation coefficient (r)
Ear density	-1.980E-02
Curvature of first rachis segment	-1.643E-02
Glaucosity of the flag leaf sheath	-1.423E-02
Spiculation of awn margins	-1.330E-02
Shape of the tip of the sterile spikelet	-1.133E-02
Length of first rachis segment	-1.065E-02
Plant length	-1.037E-02
Length of glume c.f. grain	-9.720E-03
Disposition of lodicules	-9.120E-03
Collar type	-7.890E-03
Ear attitude	-6.980E-03
Ear shape	-6.550E-03
Time of ear emergence	-5.740E-03
Frequency of plants with re-curved leaves	-5.690E-03
Ear length.	-2.390E-03
Adherence of grain husk	-1.920E-03

8.5 Association Mapping.

Evidently the test statistic of association has been considerably inflated for at least some of the traits. I propose to consider a group of traits with low inflation and a group with high inflation of the test statistic as exemplars and will summarise the remaining traits graphically in appendices containing figures reproduced from Cockram *et al.* 2010 together with additional material.

Methods for control of spurious association will be those described previously (Chapter 3) namely; Genomic Control (GC), Structured Association (SA), Correction with Principal coordinates (PC) and Mixed Effects Modelling (MM).

8.5.1 Genomic Control.

The 1 d.f. association statistic used in 8.4.1, above, was adjusted to have a median of 0.455 which is the median of a chi-squared distribution 1df under the null hypothesis.

Where inflation is modest (Figure 8-4) the effect of GC is small, however, it is quite possible for a trait exhibiting strong association with a few markers to return an association statistic which is otherwise not inflated (the anthocyanin intensity of awn tips being an example).

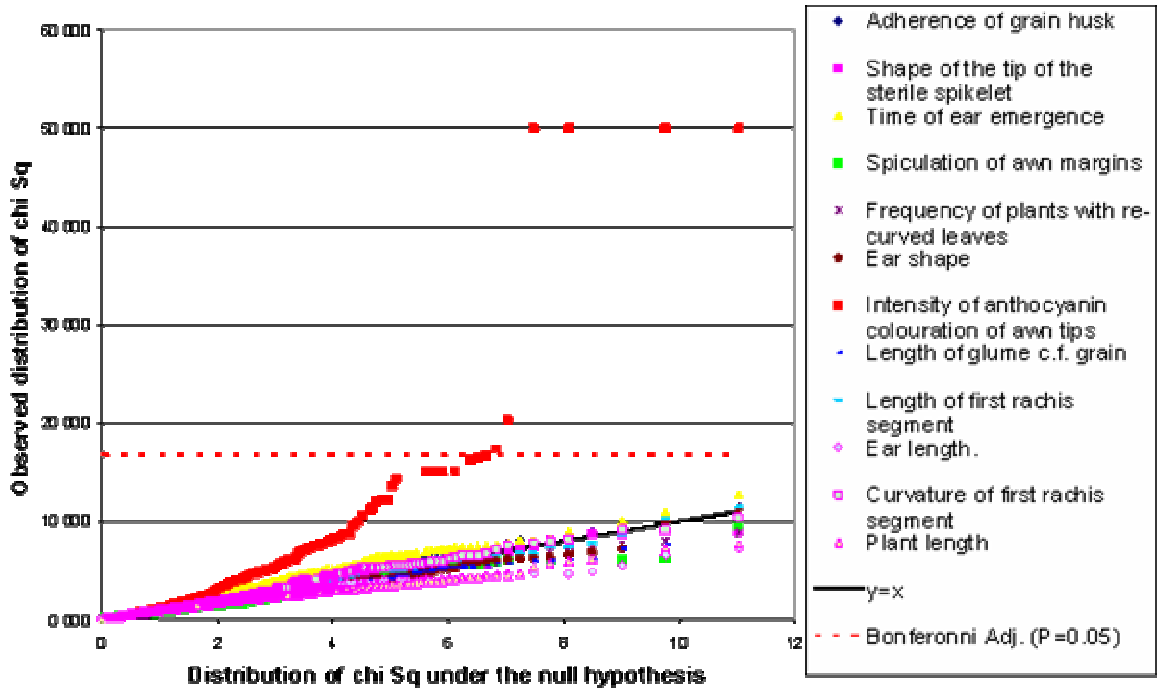


Figure 8-4. Q-Q plot for the eleven traits exhibiting little inflation of the test statistic. The empirical distribution of the statistic for these traits has been adjusted to have median 0.455. The anthocyanin intensity trait shows a group of highly associated markers which remain significant at the experiment-wise threshold of P=0.05 threshold even after this adjustment. Note: for presentation, values of chi-sq greater than 50 have been set to 50.

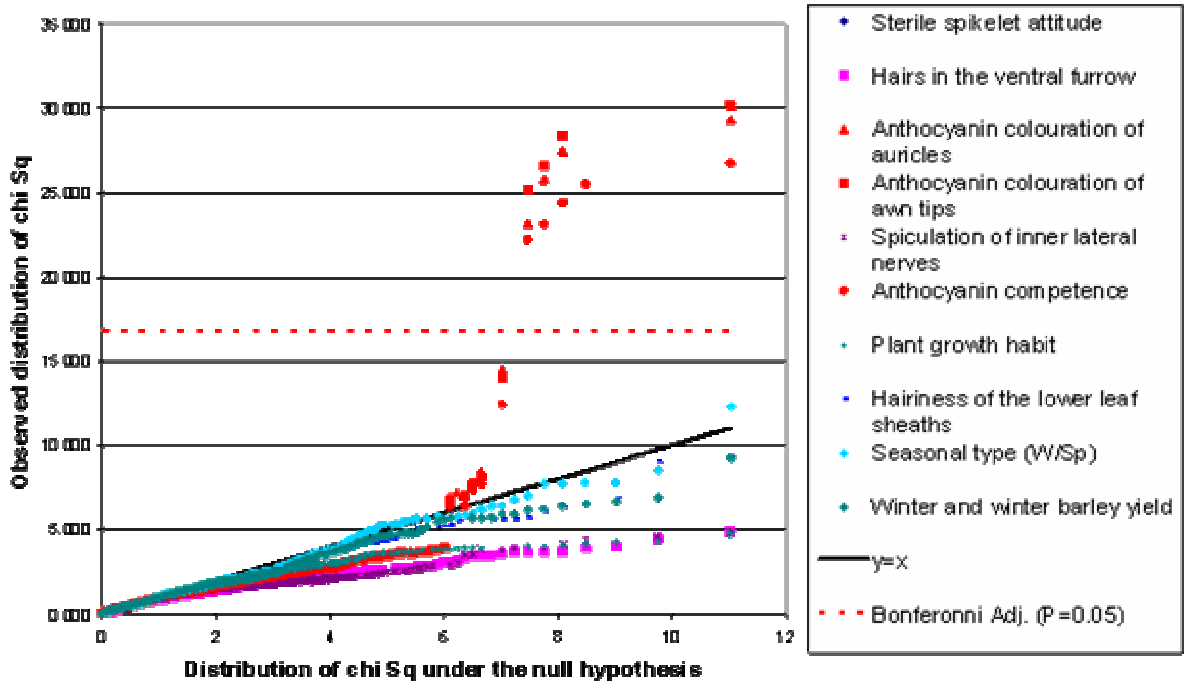


Figure 8-5. Q-Q plot for the ten traits exhibiting extreme inflation of the test statistic. The empirical distribution of the statistic for these traits has been adjusted to have median 0.455. The anthocyanin traits (red) show a group of highly associated markers which remain significant at the experiment-wise threshold of P=0.05 threshold even after this adjustment.

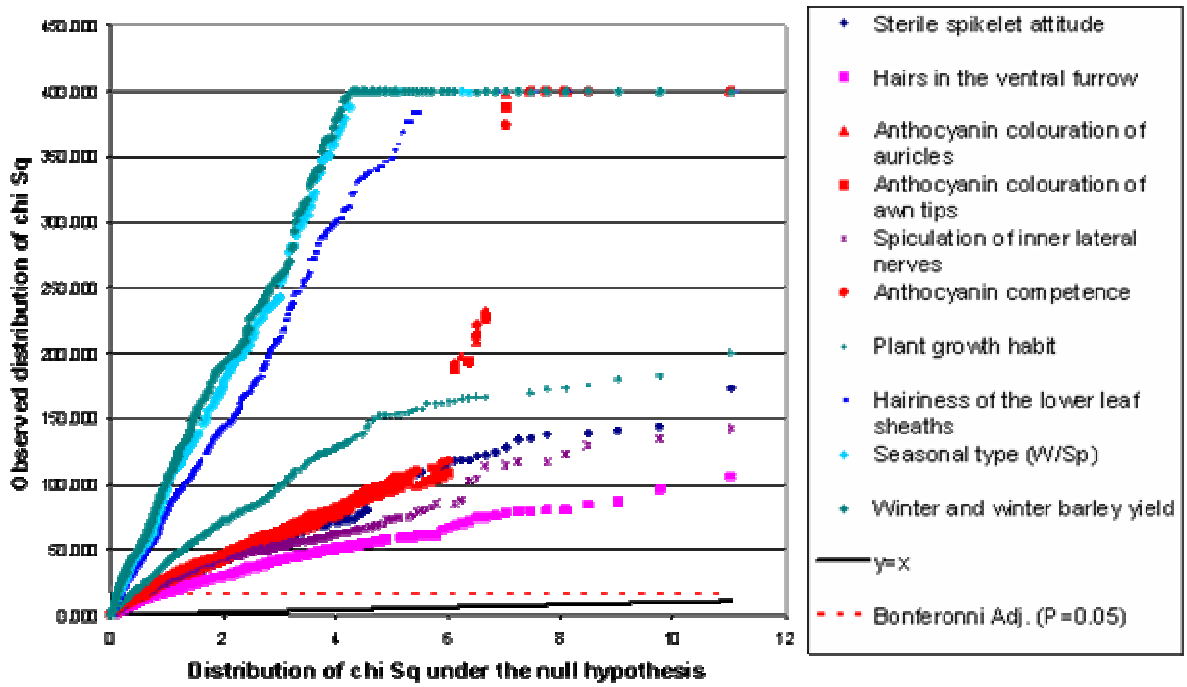


Figure 8-6. Q-Q plot for the ten traits exhibiting extreme inflation of the test statistic. Empirical values of the statistic without adjustment suggest almost every marker is significantly associated with these traits. Note: for presentation, values of chi-sq greater than 400 have been set to 400.

8.5.2 Significantly associated markers.

The application of GC deflates the test statistic and so attempts to exclude the alternate hypothesis of genome-wide inflation as an explanation for association. Multiple testing requires the use of an experiment-wise significance threshold, initially I have used a Bonferroni adjustment. However, since the marker genotypes are often correlated with each other this is likely to be excessively conservative.

The alternative adopted previously (Chapter 7) was to use a threshold based on q-value (FDR) (see; 3.5.1). In these barley data the extreme inflation in many of the traits and the relatively small marker panel means that GC results in a probability distribution skewed towards high p-values. Attempting to calculate q-values in such a distorted data-set is meaningless since the underlying null distribution is obscured. It is, however, possible to apply q-values to the unadjusted results of association. In some respects the result is analogous to the use of GC since the effect of the procedure is to adjust the significance threshold without altering the ranking of significance.

One of the advantages of using q-values as a significance threshold is the ability to objectively manage the number of discoveries; in this case the inflation of the test statistic is such that I use q=0.001 as a threshold which appears to differentiate between general inflation and clear peaks of significance, however this choice is fairly arbitrary.

Traits appear to divide into three classes; single clear peak of association (see example in Figure 8-7), general genome-wide significance (see example in Figure 8-8) and no strong evidence for association at any point (see example in Figure 8-9). Plots for all traits are presented in Appendix 7.

For reference I list the traits together with the locations of any highly associated markers in Table 8-4 (first column).

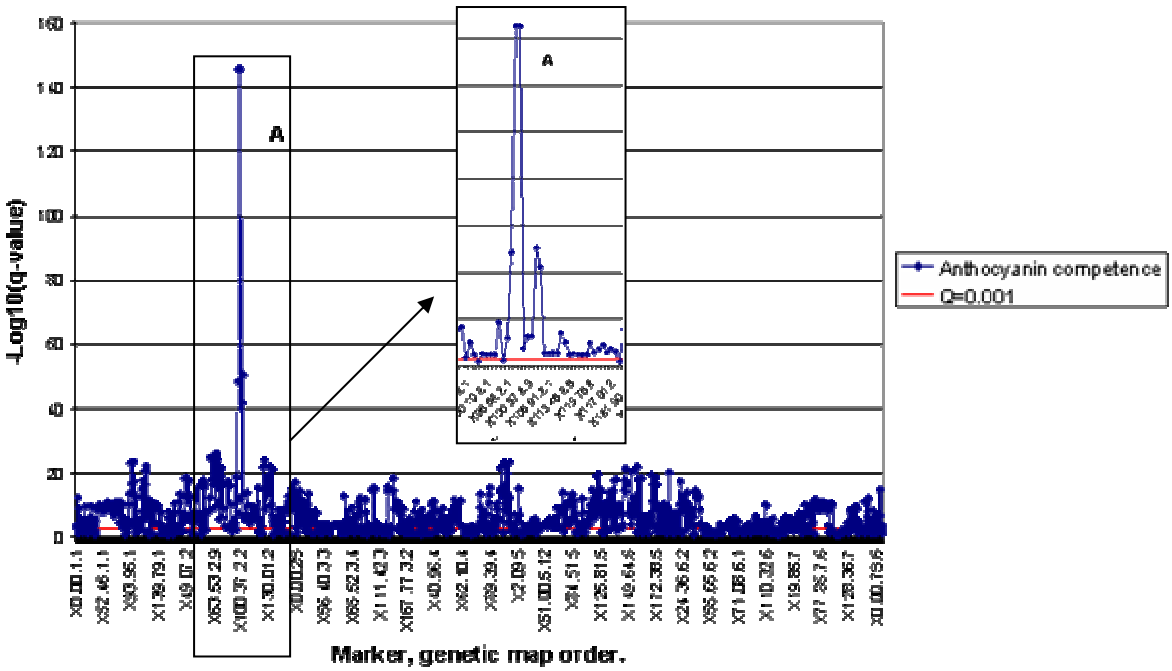


Figure 8-7. Anthocyanin competence. –log (q-value) over marker panel, markers are in genetic map order. Although there is genome wide significance at q=0.001 there appears to be a group of markers at about 96cM on Chromosome 2 (A) which are exceptionally strongly associated with the trait. This location is similarly highlighted in all the anthocyanin related traits.

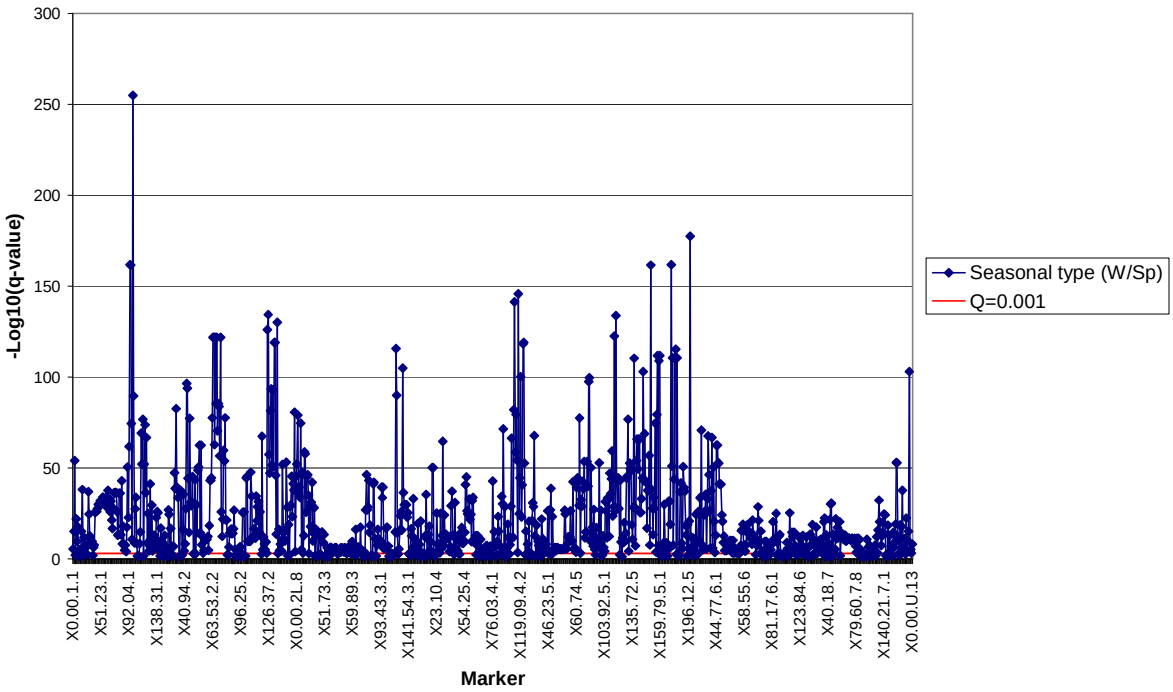


Figure 8-8. Seasonal type. Q-value over marker panel, markers are in genetic map order. There is genome wide significance at q=0.001 and no single locus appears to dominate the pattern of association.

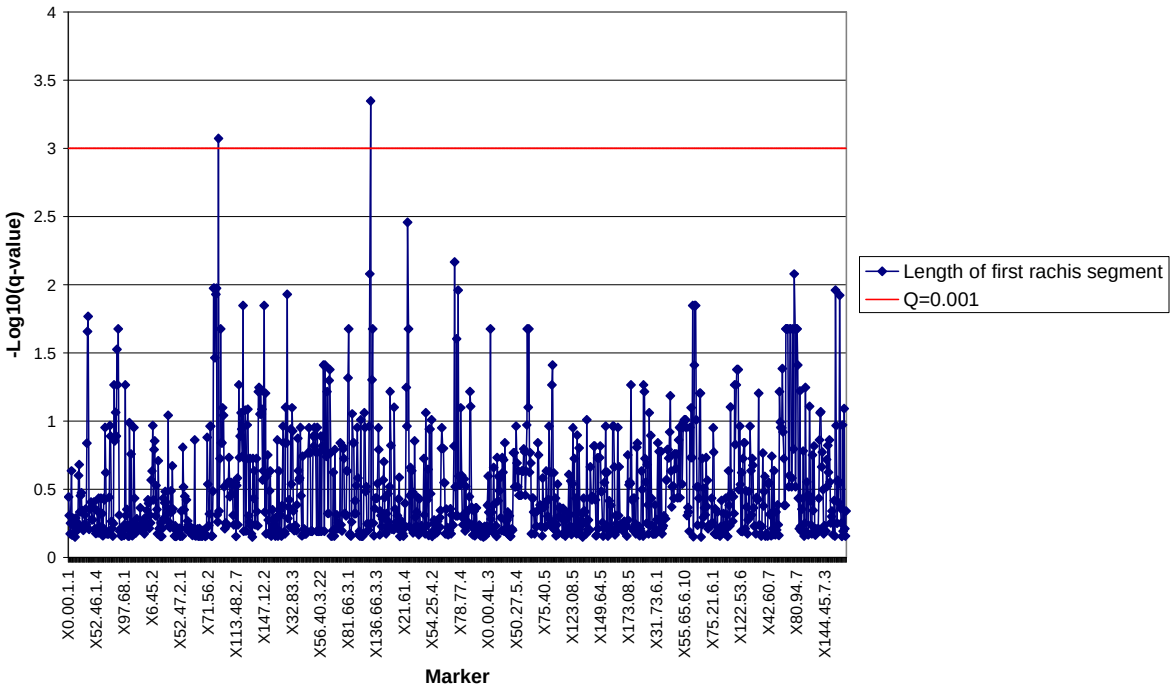


Figure 8-9. Length of the first rachis segment. Q-value over marker panel, markers are in genetic map order. There is only weak evidence for association of markers with this trait.

Table 8-4. Comparison of locations of significant markers between the four methods applied in this chapter.

Phenotype name	Location of significant peaks found by regression	Location of significant peaks found by SA	Location of significant peaks found by PCA	Location of significant peaks found by MM
Adherence of grain husk				
Shape of the tip of the sterile spikelet				
Time of ear emergence				Ch 2: ~63 cM[1] Ch 5: ~129cM[1]
Spiculation of awn margins				
Frequency of plants with re-curved leaves				
Ear shape		Ch2: ~115cM		
Intensity of anthocyanin colouration of awn tips	Ch 2: ~94 cM	Ch 2: ~94 cM	Ch 2: ~94 cM	Ch 2: ~94 cM[2] Ch 7: ~62 cM[3]
Length of glume c.f. grain				
Length of first rachis segment				
Ear length.				
Curvature of first rachis segment				Ch 1: ~116cM[1] Ch 5: ~159cM[1]
Plant length		Ch4: ~81 cM		Ch 2: ~87cM[1] Ch 4: ~51cM[1] Ch4: ~81 cM[1]
Spring barley yield		ChA: ~55 cM ChE: ~73 cM ChG: ~47 cM ChF: ~84 cM		ChC: ~82 cM[1]
Collar type				
Ear density				
Disposition of lodicules	Ch 5: ~135cM		Ch2: ~137 cM Ch4: ~51 cM Ch5:~135 cM	Ch2: ~137 cM[1] Ch4: ~51 cM[1] Ch5:~135 cM[1]
Ear attitude	Ch 5: ~135cM	Ch2: ~83 cM Ch 5: ~135cM		Ch 5: ~135cM[2]

[1] Not reported in Cockram *et al.* 2010.
[2] Major association in Cockram *et al.* 2010 (≥ 2 peaks above Bonferroni threshold within 4cM window).
[3] Minor association in Cockram *et al.* 2010 (single peak above Bonferroni threshold within 4cM window).

… continued.

Phenotype name	Location of significant peaks found by regression	Location of significant peaks found by SA	Location of significant peaks found by PCA	Location of significant peaks found by MM
Glaucosity of the flag leaf sheath		Ch2: ~58 cM	Ch2: ~58 cM	Ch2: ~39 cM[1] Ch2: ~58 cM[1] Ch5: ~130cM[1] Ch7S: ~unmapped[3]
Rachilla hair type	Ch 5: ~100cM	Ch 5: ~100cM	Ch 5: ~100cM	Ch 5: ~100cM[3]
Winter barley yield		ChC:~105 cM ChC:~137 cM ChG:~55 cM		ChC: ~85cM ChE: ~ 56cM ChG: ~26cM ChG: ~ 55cM ChF: ~117cM ChB:~112cM
Intensity of anthocyanin colouration of auricles	Ch 2: ~94 cM	Ch 2: ~94 cM	Ch 2: ~94 cM	Ch 2: ~94 cM[2]
Awn length			Ch 6: ~28 cM	Ch 6: ~28 cM[1]
Anthocyanin colouration of grain lemma nerves	Ch 2: ~94 cM	Ch 2: ~94 cM	Ch 2: ~94 cM	Ch 2: ~94 cM[2]
Colour of the aleurone layer	Ch 4: ~63 cM	Ch 4: ~63 cM	Ch 4: ~63 cM	Ch 4: ~63 cM[2]
Ear glaucosity			Ch 2: ~82 cM Ch 6: ~42 cM	Ch1: ~1cM[1] Ch1: ~66cM[1] Ch 6: ~42 cM[3]
Row number (6/2)	Ch4: ~26 cM		Ch 4: ~26 cM	Ch 1: ~75 cM[3] Ch 2: ~85 cM[3] Ch 4: ~26 cM[2] Ch 4: ~117cM[3] Ch 6: ~105cM[3]
Development of the sterile spikelet	Ch2: ~89 cM	Ch2: ~85 cM	Ch2: ~85 cM	Ch2: ~85 cM[2]
Sterile spikelet attitude		Ch1: ~54cM Ch2: ~88cM	Ch1: ~54 cM Ch2: ~88 cM	Ch1: ~54 cM[2] Ch2: ~88 cM[2]
Hairs in the ventral furrow		Ch6: ~1cM	Ch 2: 64 cM Ch6: ~1 cM Ch6: ~44 cM	Ch 4: ~26 cM[1] Ch6: ~1 cM[2]
Anthocyanin colouration of auricles	Ch 2: ~94 cM	Ch 2: ~94 cM	Ch 2: ~94 cM	Ch1: ~75cM[3] Ch 2: ~94 cM[2]
Anthocyanin colouration of awn tips	Ch 2: ~94 cM	Ch 2: ~94 cM	Ch 2: ~94 cM	Ch1: ~75cM[3] Ch 2: ~94 cM[2]

[1] Not reported in Cockram *et al.* 2010.

[2] Major association in Cockram *et al.* 2010 (≥ 2 peaks above Bonferroni threshold within 4cM window).

[3] Minor association in Cockram *et al.* 2010 (single peak above Bonferroni threshold within 4cM window).

… continued.

Phenotype name	Location of significant peaks found by regression	Location of significant peaks found by SA	Location of significant peaks found by PCA	Location of significant peaks found by MM
Spiculation of inner lateral nerves		Ch2: ~52-87cM	Ch2: ~78cM	Ch4: ~26cM[1] Ch4: ~54cM [1] Ch2: ~78cM [2] Ch2: ~151cM[1] Ch5: ~180cM[1]
Anthocyanin competence	Ch 2: ~94 cM	Ch 2: ~94 cM	Ch 2: ~94 cM	Ch 2: ~94 cM
Plant growth habit				Ch 1: ~96cM[3]
Hairiness of the lower leaf sheaths		Ch4: ~111 cM	Ch4: ~111 cM	Ch1: ~96cM[3] Ch2: ~126cM[3] Ch4: ~111cM [2] Ch5: ~153cM[3]
Seasonal type (W/Sp)		Ch5: ~137 cM	Ch 1: ~96cM Ch 5: ~137cM	Ch 1: ~96cM[2] Ch 5: ~100-161cM[3]
Winter and spring barley yield			Ch A: ~96cM ChG: ~117cM	Ch A: ~96cM Ch C: ~50-82cM Ch G: ~117cM Ch F: ~150-170cM Ch D: ~105cM

[1] Not reported in Cockram *et al.* 2010.

[2] Major association in Cockram *et al.* 2010 (≥ 2 peaks above Bonferroni threshold within 4cM window).

[3] Minor association in Cockram *et al.* 2010 (single peak above Bonferroni threshold within 4cM window).

In this association analysis it is evident that some traits can already be said to have clear candidate locations for trait loci, others show association over the entire genome and some traits, notably yield, show no association with any markers in this analysis.

Attempting to incorporate models for structure into the regression should serve to confirm that the peaks of association noted are not artefacts. Additionally traits showing generalised association may reveal candidate QTL in the more sophisticated model. Finally, it is possible that more sophisticated modelling will reveal associations even in the traits currently showing no significant markers whatsoever.

8.6 Structured Association.

As in the wheat case the first task is to determine the appropriate value for *k.* To do this a sub-set of independent markers was chosen separated by about 2cM this amounted to 307 loci. All available varieties were included in the model. STRUCTURE was run with a burnin of 250,000 and MCMC of 1,000,000 cycles for *k*= 1 to 19. Replicate runs were conducted until pairs were obtained for which the average maximum correlation between the columns of the resulting **Q** matrices was 0.999 or greater.

The appropriate value for *k* was determined by considering the rise in the likelihood of the model (**Figure 8-10**), Δk (**Figure 8-11**) and ΔF_{ST} (**Figure 8-12**). Although ln(P(D)) rises gently until *k*=19 there is a reduced gradient above *k*=7, the peaks of Δk fall at k= 2, 6, 10 and 17 the largest being k=2 but at this point ln(P(D)) is still sising rapidly ΔF_{ST} peaks at k=2 and slightly at k=10. The major structural features of barley are the four season/row number combinations but there is no strong evidence that k=4 is a particularly stable model.

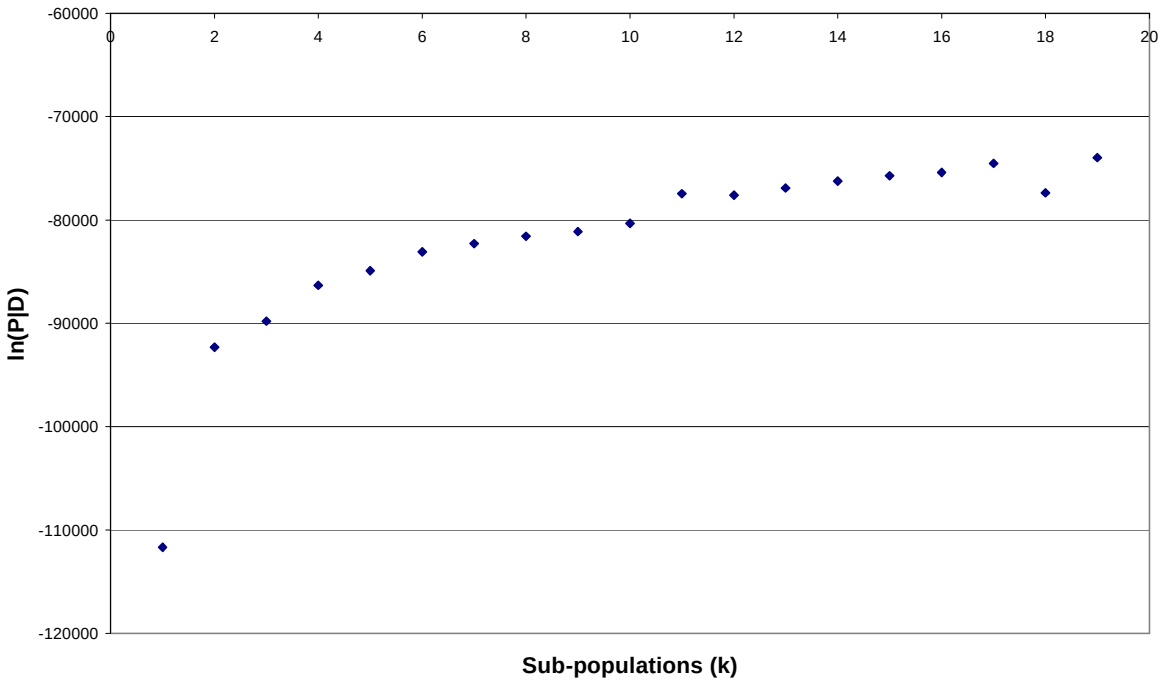


Figure 8-10. The ln(P(D)) of models of structure determined at increasing values of *k* (sub-population number). Values are the mean of two matching (highly correlated) replicate simulations.

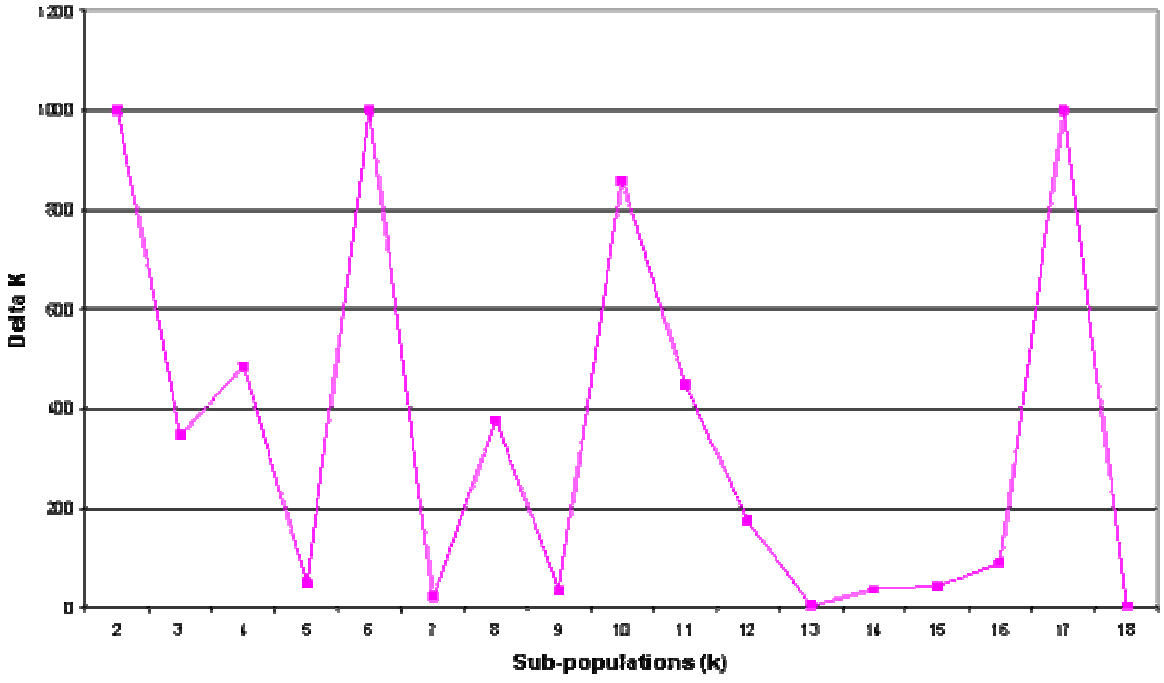


Figure 8-11. The Δ*k* statistic plotted over sub-population number. Values of Δ*k* greater than 1000 have been set to 1000 for illustration purposes. The largest value (119,000) was observed for *k* = 2. The value at *k*=6 is 7400.

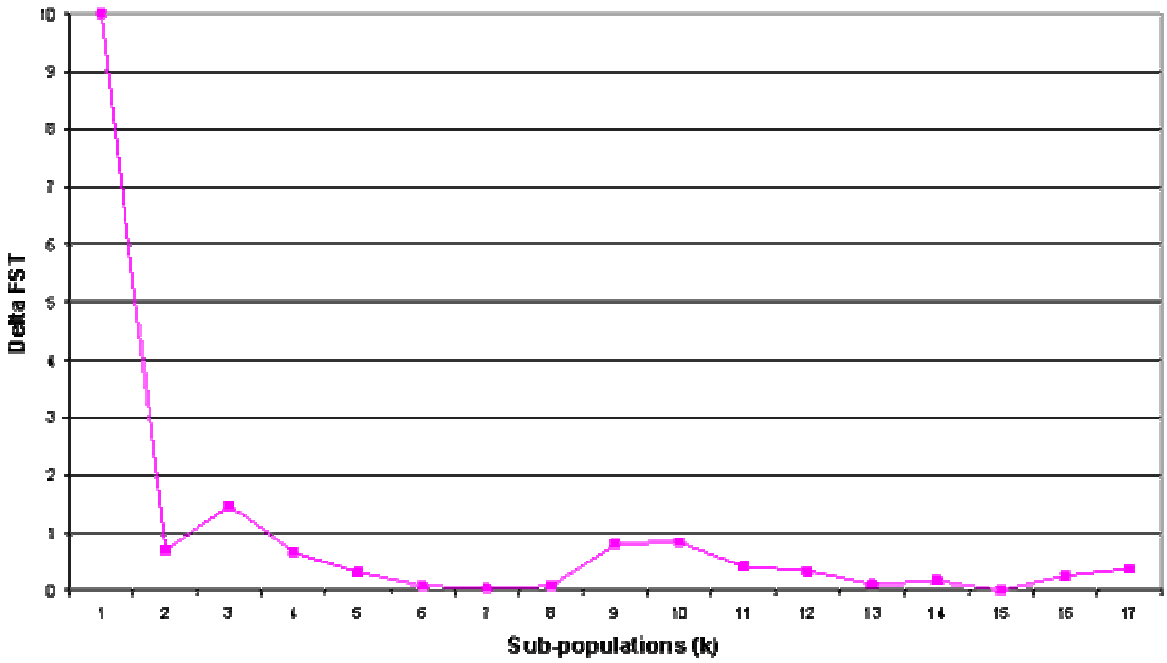


Figure 8-12. The ΔF_{ST} statistic plotted over sub-population number. Values of ΔF_{ST} greater than 10have been set to 10 for illustration purposes. The largest value (37) was observed for k = 2.

Given the large number of phenotypes to test I chose to use a single Q matrix defined at k=10 because it exceeded the value of k necessary to define the known major structural components of barley, it corresponded to a plateau in ln(P(D)) over k and was coincident with peaks for the two other metrics.

Using a Q matrix as a model for structure should reduce the genome-wide inflation of the test statistic in a selective manner; marker genotypes which co-vary with structure will offer no additional information about phenotypic variation if they are added to the model after the structure terms. Markers which co-vary with both trait and structure may explain additional variance when added to the model.

The model used for structured association is therefore an analysis of variance of a general linear model (with a logistic link function). If the binary genotype is predicted using the first (k-1) columns of **Q** and subsequently the phenotype then the change in deviance following the addition of phenotype is a measure of association of genotype and phenotype after correction of the structure model. In this case the change in deviance approximates to chi-sq 1df.

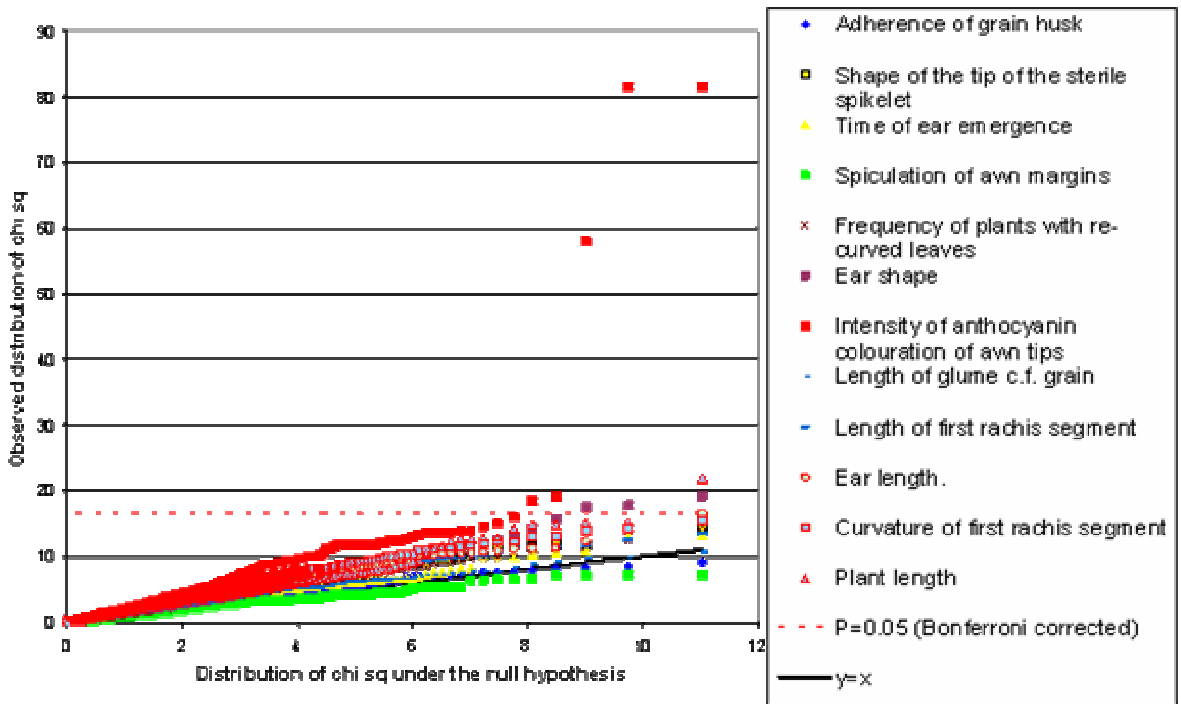


Figure 8-13. Q-Q plot. Structured Association. The traits which exhibited low inflation of the test statistic under simple regression also show low inflation following SA. The anthocyanin trait shows extreme statistical significance of association in three markers.

Figure 8-13 shows significant association (P=0.05; experiment-wise) for some traits, notably the intensity of anthocyanin colouration of awn tips. Comparison with Figure 8-4 reveals that the SA adjustment has substantially reduced the estimate of statistical significance of the most significant markers such that the anthocyanin trait has only three cases of extreme significance (compared to four previously). Additionally, the traits showing weak, but significant, association change between the two figures; in Figure 8-4 length of first rachis segment and curvature of first rachis segment showed some evidence of associated markers, following SA these have been replaced by plant length and ear length. Evidently the regression model is capable of selective deflation and inflation of the significance of markers.

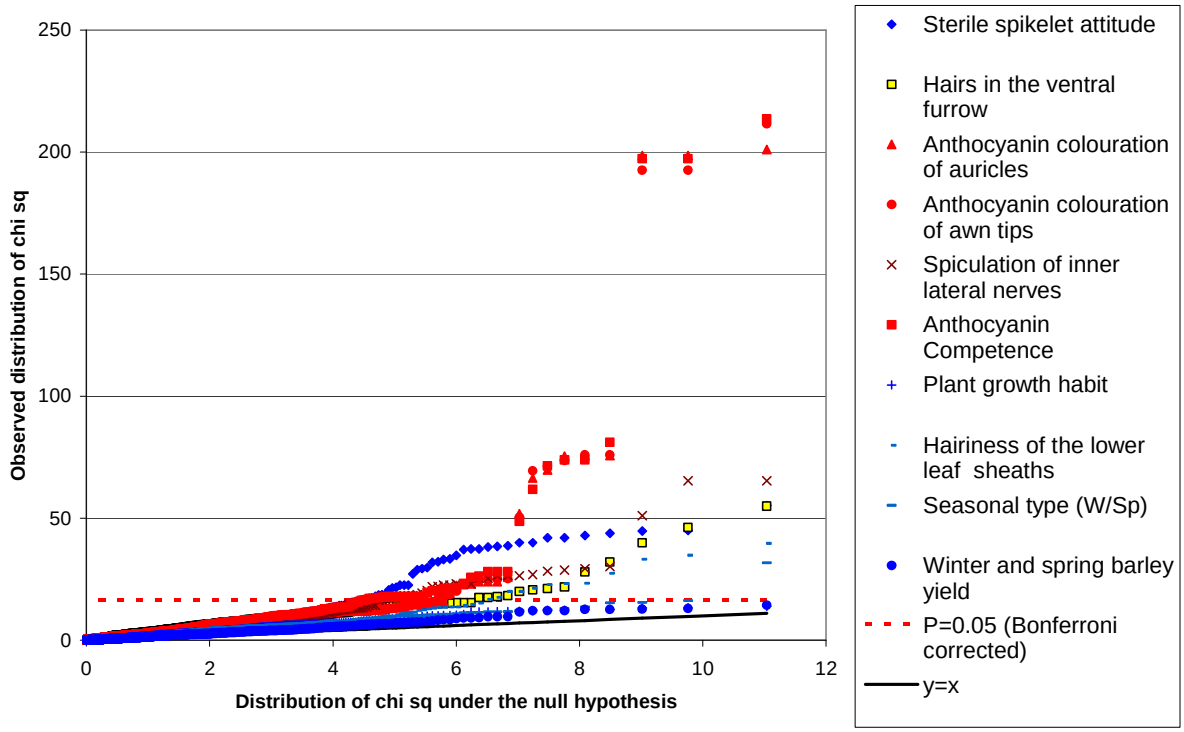


Figure 8-14. Q-Q plot. Structured Association. The traits which exhibited extreme inflation of the test statistic under simple regression show reduced inflation following SA, however there is enough residual inflation to make a substantial proportion of markers appear significantly associated with most traits.

Considering Figure 8-14, the traits showing extreme inflation of the test statistic under regression, comparison with Figure 8-6 shows that SA has deflated the test statistic but the number of apparently associated markers for the majority of these traits still seems excessive and there remains a general tendency for the Q-Q plot to deviate from y=x. It appears that the regression model for structure is sufficiently incomplete to allow an important amount of distortion to persist. As previously (7.3) an additional adjustment with GC can be used to remove this effect.

Figure 8-15 shows the effect of the additional adjustment with GC. In this figure six traits (sterile spikelet attitude, anthocyanin colouration (auricles, awns and competence), spiculation of the ILN, hairs on the lower leaf sheath and seasonal type) exhibit significant association with at least one marker plant growth habit, yield of winter and spring barley, seasonal type and hairiness of the ventral furrow show no significant association with any marker. Comparison with Figure 8-5 reveals that the SA+GC model has identified markers associated with spiculation of the inner lateral nerves which were not found by simple regression. In addition, SA has not found the association with winter and spring yield or seasonal type. Thus the more sophisticated regression model has selectively promoted and demoted marker associations.

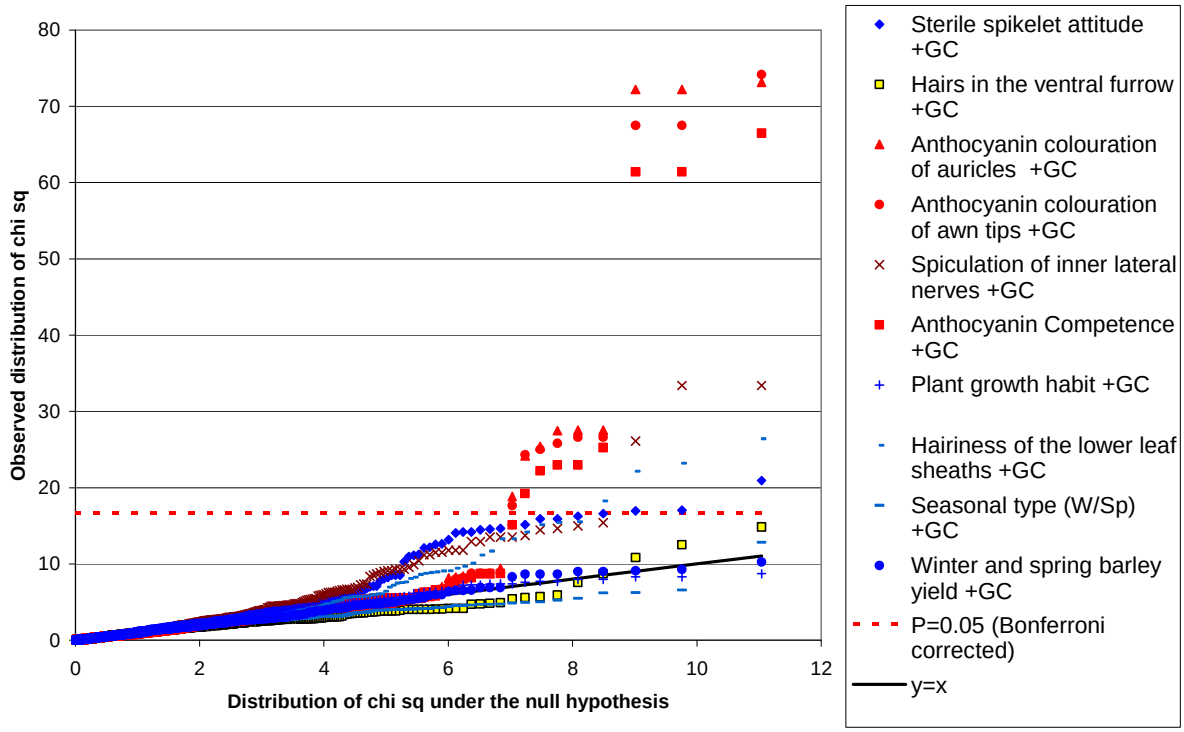


Figure 8-15. Q-Q plot. Structured Association. The traits which exhibited extreme inflation of the test statistic under simple regression showed reduced inflation following SA. Here an additional adjustment using genomic control has been used to remove this generalised inflation. Six of the traits still exhibit significant association with at least one marker.

8.6.1 Significantly associated markers.

Following SA the distortion of the underlying null distribution seen in simple regression is largely absent so it is possible to calculate q-values for the SA+GC treatment and to use these as an objective measure of significance. The deflation of the test statistic renders many fewer markers potentially significant and the threshold for significance must be raised to q=0.1 to allow a useful number of discoveries. It may seem that adjusting thresholds between methods makes between method comparison meaningless but the data are showing clear peaks of association; the adjustment of threshold allows differentiation between these peaks and the background noise. Critically, the more sophisticated regression model has changed the composition of the class ‘discoveries’. This has been done by removing discoveries made under the alternate hypothesis i.e. that ‘trait-marker association reflects trait association with population sub-structure’. Thus absolute significance is lower but power to differentiate null from alternate is higher.

When using simple regression the genome-wide inflation resulted in general association over the whole genome for many traits. Determining significance with q-values revealed three classes of trait; those with no significant association, those with single peaks of association and those which retained widespread association over the genome.

Following SA and GC the pattern of association over traits is different; many traits exhibit no significant association (e.g. Figure 8-16) and some exhibit one or more peaks of association (e.g. Figure 8-17). The general genome-wide associations seen previously are no longer evident (e.g. Figure 8-18). What is perhaps most interesting is the contrast between the association patterns seen following analysis with the two models, I summarise this in Table 8-4.

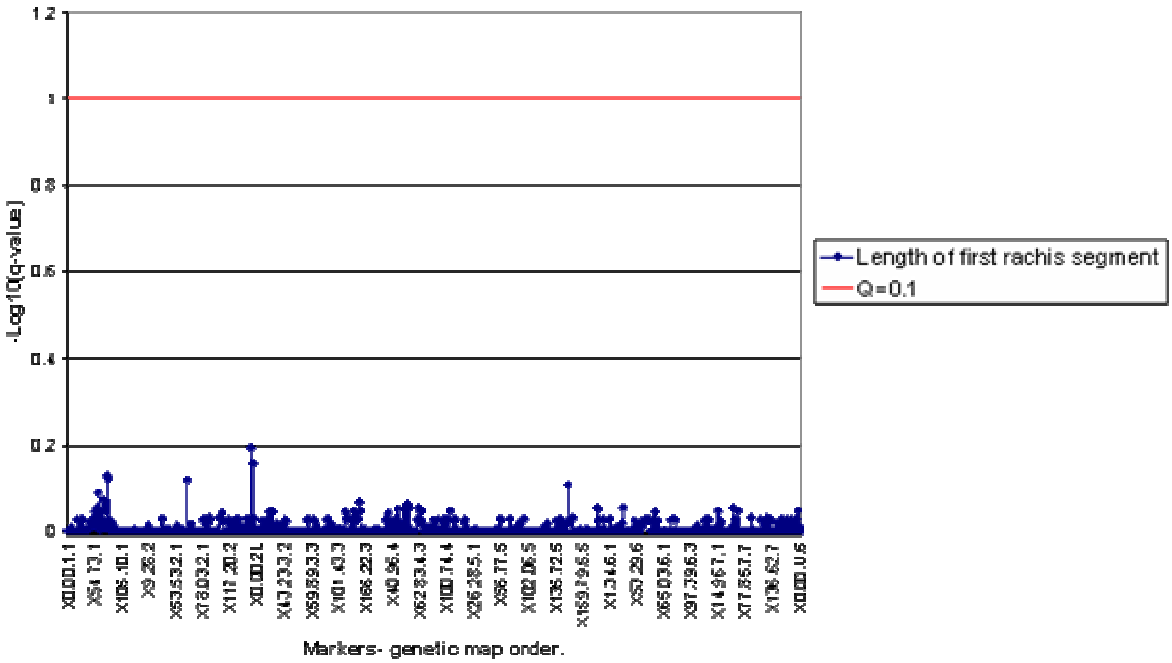


Figure 8-16. Length of the first rachis segment. Under simple regression association with the trait was comparatively weak and uniform over the genome. Following SA this result is confirmed, there is no evidence for marker-trait association in these results.

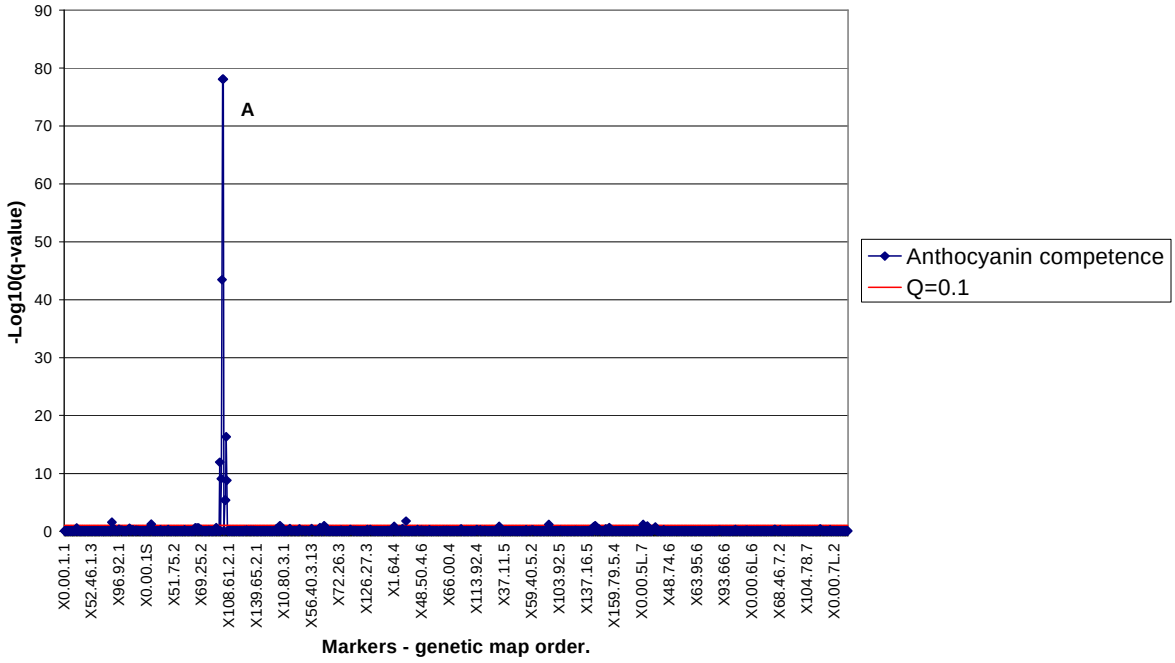


Figure 8-17. Anthocyanin competence. Q-value over marker panel, markers are in genetic map order. The single peak of association seen in simple regression analysis persists (96cM on Chromosome 2 (A)); the generalised genome-wide association has been completely controlled.

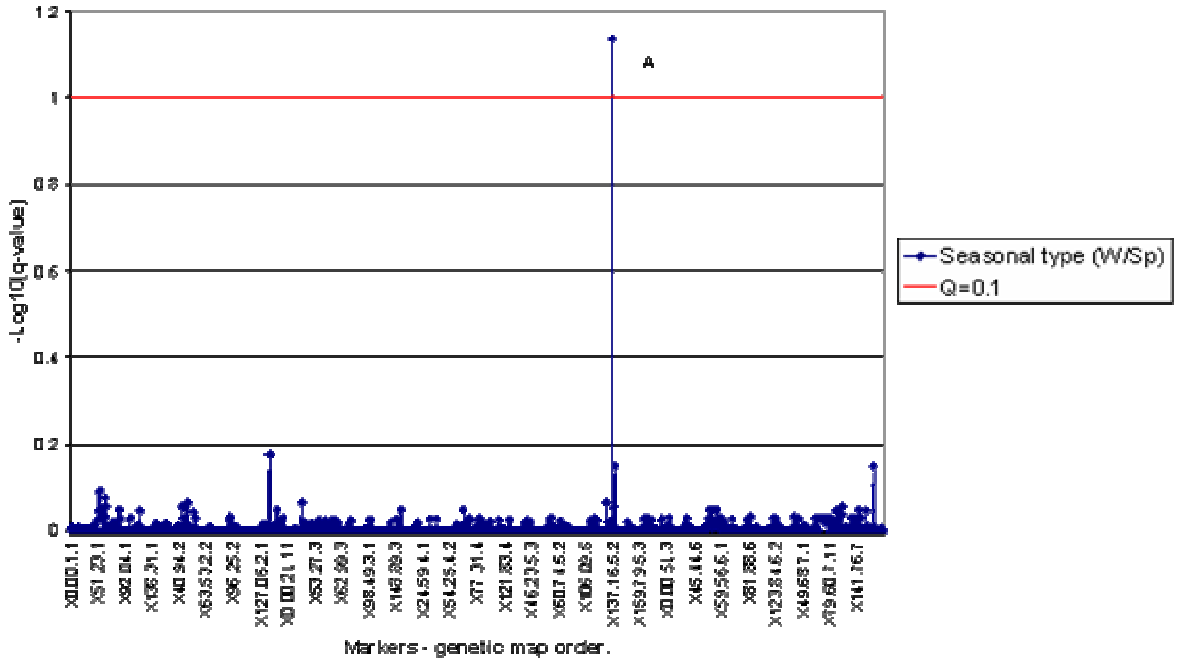


Figure 8-18. Seasonal type. Q-value over marker panel, markers are in genetic map order. The genome-wide association seen in simple regression has been removed; one marker appears significant at Q=0.1. The position of the marker A (Ch5: ~137 cM) is within 5cM of VRN-1, a gene responsible for vernalisation response, the physiological mechanism behind 'seasonal type' it is also within a cluster of QTL linked to frost tolerance.

Table 8-4shows that of the 12 trait associations found by simple regression, disposition of lodicules and row number (6/2), were not confirmed following SA+GC. Conversely, the application of SA+GC removed all cases where there was significant genome-wide association and from within the genome-wide association SA+GC found discrete significant associations with spiculation of inner lateral nerves, hairiness of the lower leaf sheath and seasonal type (W/Sp). In addition, ear shape, plant length, spring barley yield, glaucosity of the flag leaf sheath and winter barley yield, which had all exhibited no clear association under the simple regression model were found to have significantly associated markers using SA. Plots of -log q-value over markers for all traits following SA+GC are given in Appendix 8.

The structure model achieves marker-specific adjustment of association and results in fewer markers being statistically significantly associated with trait but those markers which are associated are more likely to be so due to physical proximity of QTL and marker in the genome because a large class of discoveries attributable to structure have been excluded.

8.7 Correction using principal coordinates.

Modelling structure as principal coordinates of a between-varieties genetic covariance matrix and incorporating the information in a regression model should give similar power to selectively deflate and inflate individual marker-trait associations as demonstrated with SA (8.6). However, calculating the eigenvectors necessary is a far less onerous task than calculating and evaluating the characteristics of a range of **Q** matrices.

The model used for testing association comprises simple regression between the residuals of phenotype and genotype separately predicted using the top few eigenvectors of the covariance matrix. In this case the top 20 vectors have been used for all phenotypes, this is the default recommended by Price *et al.* (2006). Significance can be estimated from an analysis of variance or from the correlation coefficient calculated between the two sets of residuals; in the latter case the significance test is two-tailed since positive or negative associations are equally important.

If GC is to be applied the results must be expressed in terms of a 1d.f. test statistic. Where necessary in the subsequent analysis a chi squared (1df) distribution has been recovered from p-values of association from the above tests using the ‘qchisq’ function in R.

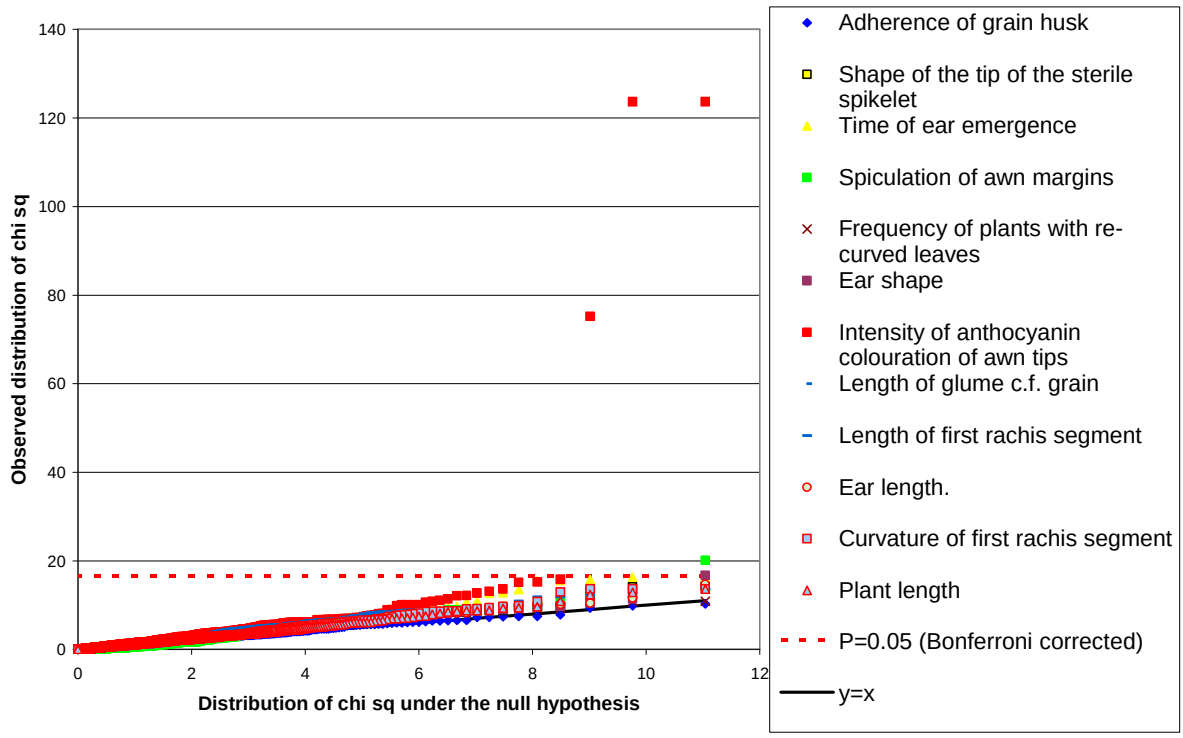


Figure 8-19. Q-Q plot. PC correction. The traits which exhibited low inflation of the test statistic under simple regression also show low inflation following PC correction. The anthocyanin trait shows extreme statistical significance of association in three markers.

Figure 8-19 shows that for traits exhibiting low inflation in simple regression analysis the PC correction has removed almost all trace of genome-wide inflation such that the observed distributions approximate to the expected over much of the range. As with SA three markers are identified which associate strongly with anthocyanin competence. The only other marker significant at the Bonferroni adjusted P=0.05 threshold associates with spiculation of awn margins. It is interesting to note that this association is not seen with SA (Figure 8-13) or GC (Figure 8-4); this type of minor inconsistency is to be expected since the two models for structure will differ in fine detail.

Considering the traits exhibiting extreme inflation, Figure 8-20, PC correction has found the same extreme association with the anthocyanin traits as seen with SA, however ‘seasonal type’ shows considerably more significantly associated markers under PC than under SA. The only trait showing no significant association is ‘plant growth habit’ which showed no

association under SA either. In contrast ‘winter and spring yield’, which showed no association under SA exhibits significant markers under PC – although these may simply be a symptom of uncontrolled inflation.

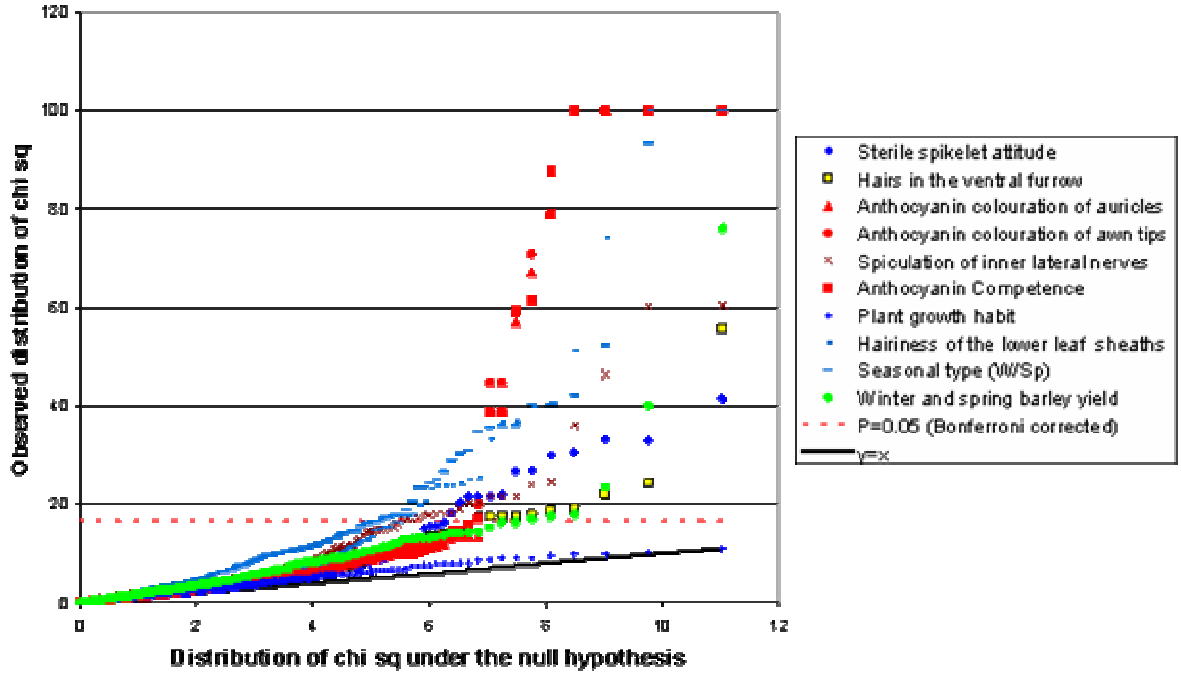


Figure 8-20. Q-Q plot. PC correction (note the y-axis has been truncated at chi sq = 100 to show detail). The traits which exhibited extreme inflation of the test statistic under simple regression show a general reduction in inflation following PC correction. However, the majority of traits are exhibiting a systematic positive deviation from y=x suggesting the control of inflation has been incomplete and further adjustment with GC might be useful.

As before, residual inflation can be controlled by GC, the results for the high inflation traits are presented in Figure 8-21. All traits except ‘plant growth habit’ retain two or more significantly associated markers.

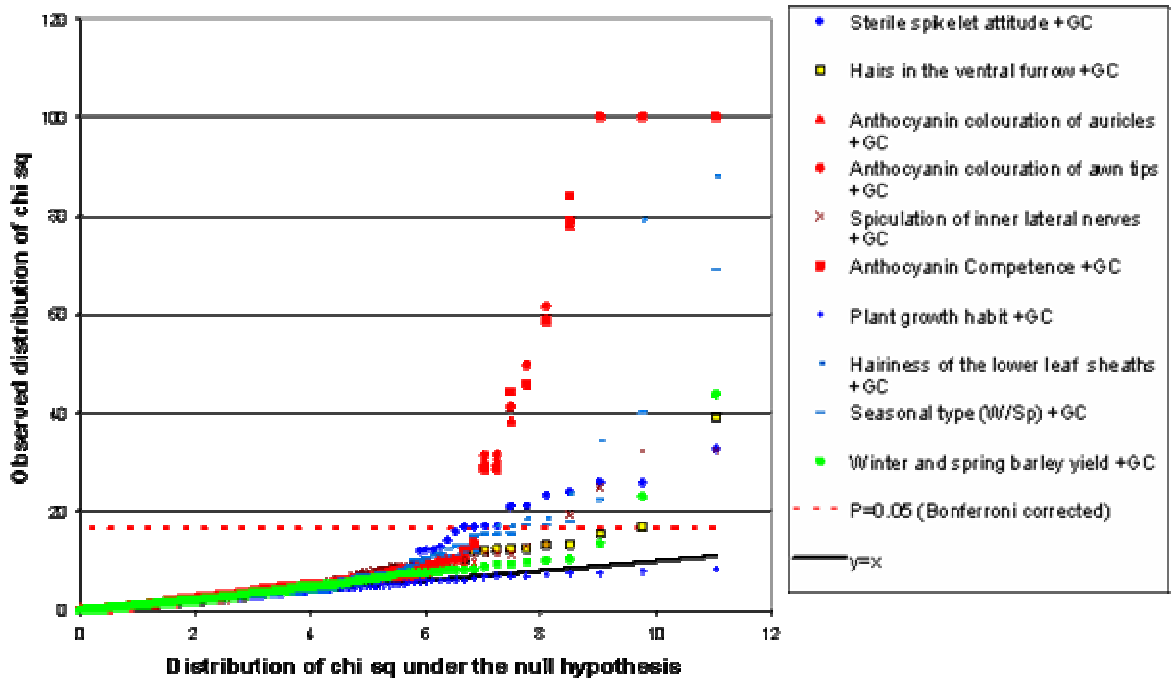


Figure 8-21. Q-Q plot. PC correction followed by an additional Genomic Control (note the y-axis has been truncated at chi sq = 100 to show detail). Although the number of significant markers for each trait is reduced by this treatment all traits exhibiting association under PC correction retain some significant markers following PC +GC.

8.7.1 Significantly associated markers.

Q-values may be used to give an experiment specific significance threshold allowing comparison between discoveries under the preceding models.

Consider the three example traits. (1) Figure 8-22 shows that PC correction reveals no markers associated with the length of the first rachis segment. (2) Figure 8-23 shows markers previously identified as associated with anthocyanin competence are again highlighted. The noise in the p-values for association for this trait using PC + GC is greater than in the case of SA +GC, consequently a number of markers elsewhere in the genome just achieve –log (q-value) greater than 1. However, the clarity of the main peak would tend suggest that the major genetic influence over the trait, in this panel, is on chromosome 2. (3) Figure 8-24 shows a contrast with the corresponding figure for SA+GC (Figure 8-18) in that the single peak on chromosome 5 close to a vernalisation gene is replaced by a cluster of markers in the same position. Additionally the peak in Figure 8-8 on chromosome 1 (96cM) which was initially dismissed as indistinguishable from general inflation is now a clearly defined peak of association. This peak is within 10cM of a known gene controlling dormancy; dormancy refers to a period post-harvest during which seeds will not germinate. Winter barley tends to

be more often dormant than does spring so this peak of association may be spurious – in the sense that it does not relate to vernalisation.

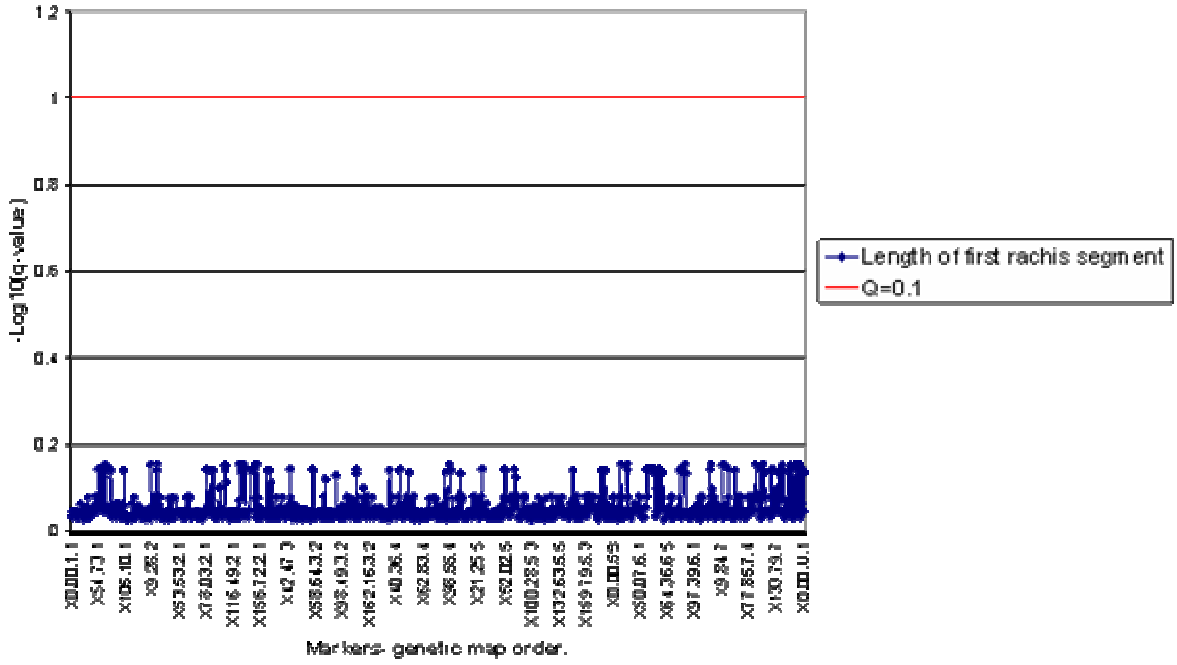


Figure 8-22. Length of the first rachis segment. Under simple regression association with the trait was comparatively weak and uniform over the genome. Following PC+GC this result is confirmed, there is no evidence for marker-trait association in these results.

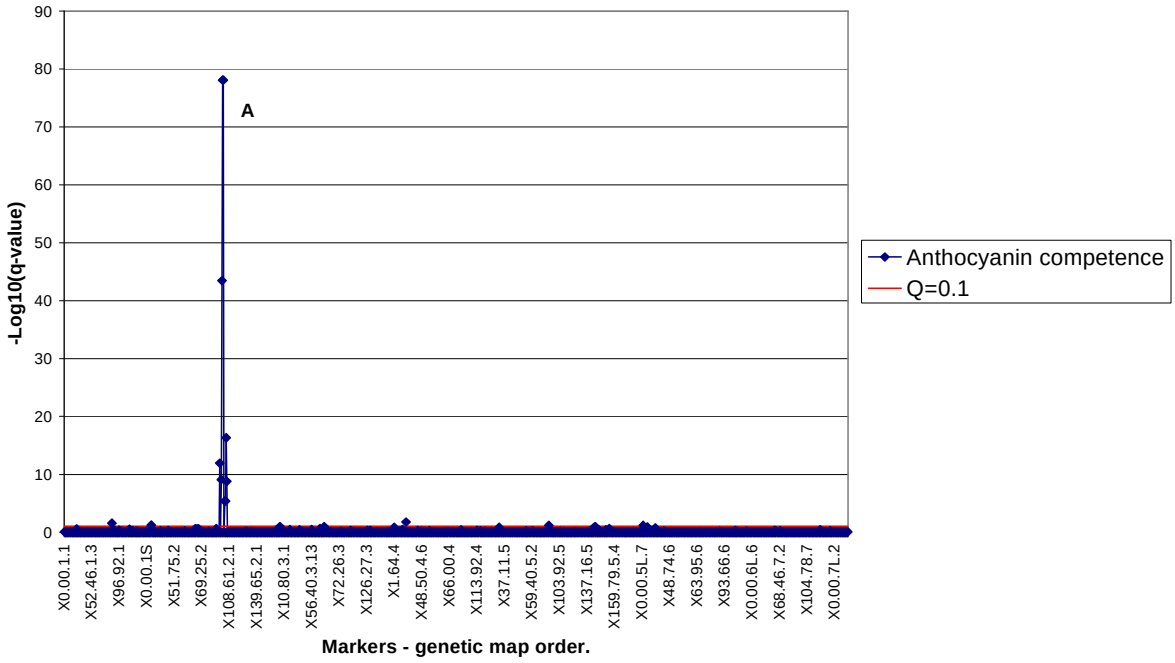


Figure 8-23. Anthocyanin competence. Q-value over marker panel, markers are in genetic map order. The single peak of association seen in simple regression analysis persists (96cM on Chromosome 2 (A));; the generalised genome-wide association has been largely controlled.

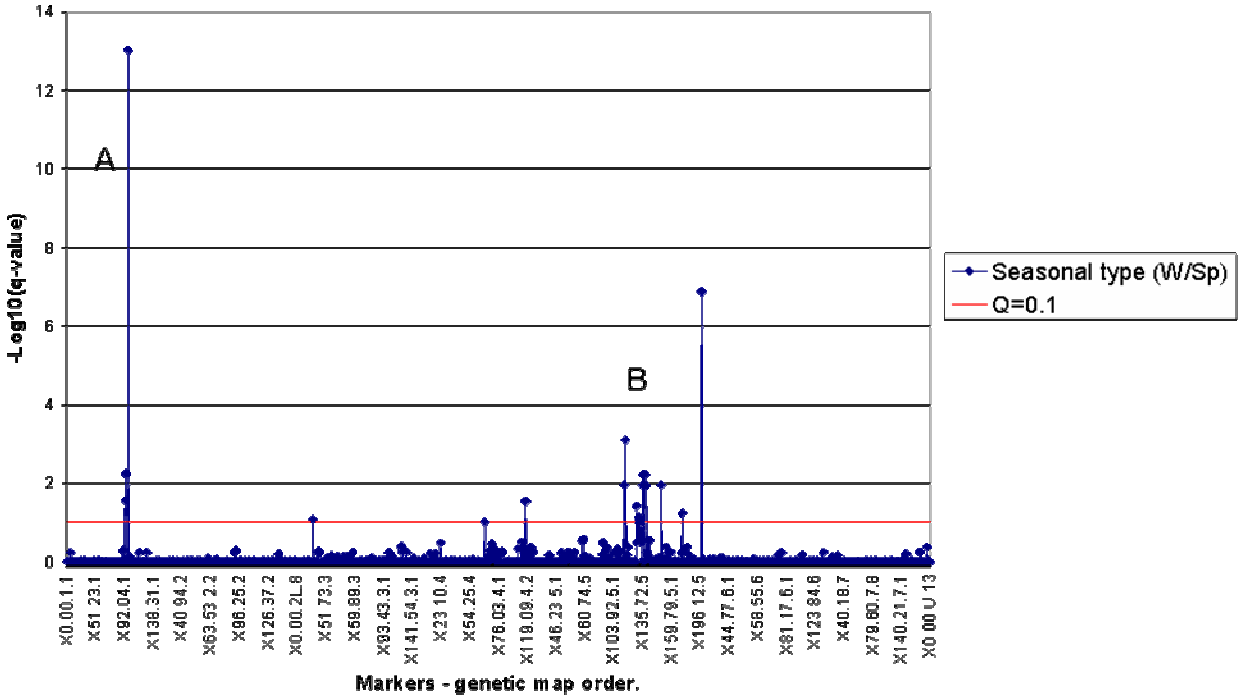


Figure 8-24. Seasonal type. Q-value over marker panel, markers are in genetic map order. The genome-wide association seen in simple regression has been removed; two regions appear to contain significant associations; (A) on chromosome 1 (96cM) the peak is within 10cm of a known dormancy QTL and (B) on chromosome 5 (100-150cM). The position of the markers on chromosome five straddle that of VRN-1, a gene responsible for vernalisation response; the physiological mechanism behind 'seasonal type'. (B) is also within a cluster of QTL linked to frost tolerance.

Association traces for all traits are presented in Appendix 9. A summary of associations significant at Q<0.1 are given in Table 8-4.

8.8 Correction using mixed effects modelling.

The use of mixed effects modelling allows the complex inter-relationships within the variety panel which can be captured in a co-ancestry or kinship matrix (**K**) to be used to account for some of the phenotypic variance. The model used is:

$$y = X\beta + Zu + e$$

Where, y = phenotype; **X** is a matrix of covariates, including genotype; β is a vector of regression coefficients (equivalent to effect sizes for these binary genotypes); **Z** is an incidence matrix linking phenotype to variety – in the model used here it is the identity matrix and u is defined such that: $Var(u) = \sigma_g^2 K$

Thus u is an n x 1 vector of the effect on phenotype predicted from **K.**

The solution to the model is found using maximum likelihoods and significance may be estimated from the change in deviance as each marker in turn is omitted from the full model; a likelihood ratio test.

In this case the model for kinship was a correlation matrix calculated using standardised genotypes (i.e. adjusted to have allele frequency = 0 and unit variance). The mixed model solution was found using the Efficient Mixed Model Analysis (EMMA; Kang *et al.* 2008) in R.

The correlation matrix used contains both negative and positive correlations; strong positive correlation suggests recent co-lineage whereas strong negative correlation suggests less than average relatedness; as argued previously this matrix contains a representation of both structure (less than average relatedness) and kinship and should offer better control of inflation than the 'structure-only' models so far deployed. Indeed, some evidence for this was seen in Figure 8-3 where the correlation between phenotypic distance and kinship was shown to be correlated with inflation of the test statistic and the correlation between phenotypic distance and kinship was consistently low when inflation was low – this contrasted with Figure 8-1 and Figure 8-2 where the correlation from structure models varied considerably at low values for inflation.

Table 8-5 summarises the inflation factors observed for all regression models, the lowest average inflation and lowest coefficient of variation was achieved with the mixed model. Indeed, since the maximum inflation factor observed with the mixed model was 1.7 I have

not used GC to further deflate the data; if GC is applied it has almost no effect on the q-values.

Table 8-5 Inflation of the test statistic observed for each trait using each regression model. Inflation is of the observed median of a chi-sq 1df statistic relative to its expectation under the null over all markers. A value of 1 indicates no inflation.

Trait	Simple	SA	PCA	MM
Winter and winter barley yield	104.3	1.4	1.7	1.3
Seasonal type (W/Sp)	95.7	2.5	2.3	1.4
Hairiness of the lower leaf sheaths	80.9	1.5	2.2	1.0
Plant growth habit	42.5	1.6	1.3	1.2
Anthocyanin Competence	30.1	3.2	1.3	1.0
Spiculation of inner lateral nerves	29.7	1.9	1.9	1.1
Anthocyanin colouration of awn tips	27.8	2.8	1.4	0.9
Anthocyanin colouration of auricles	27.4	2.7	1.4	0.8
Hairs in the ventral furrow	21.8	3.7	1.4	1.0
Sterile spikelet attitude	20.7	2.6	1.2	0.8
Development of the sterile spikelet	20.5	3.5	2.0	0.8
Row number (6/2)	19.9	1.6	2.0	1.7
Ear glaucosity	16.8	2.2	1.5	1.2
Colour of the aleurone layer	14.8	2.1	1.3	0.7
Anthocyanin colouration of grain lemma nerves	13.3	2.5	1.5	1.0
Awn length	7.9	2.1	1.1	0.9
Intensity of anthocyanin colouration of auricles	6.1	2.3	1.3	1.1
Winter barley yield	5.3	1.4	1.5	1.5
Rachilla hair type	4.6	2.4	1.7	0.9
Glaucosity of the flag leaf sheath	3.7	1.4	1.3	1.3
Ear attitude	3.1	1.5	1.4	1.2
Disposition of lodicules	3.1	1.5	1.4	1.0
Ear density	3.0	1.9	1.2	1.1
Collar type	2.9	1.4	1.3	1.4
Spring barley yield	2.7	1.7	1.4	1.0
Plant length	2.7	1.6	1.5	1.2
Curvature of first rachis segment	2.4	2.0	1.2	1.2
Ear length.	2.3	1.4	1.3	1.3
Length of first rachis segment	2.2	2.0	1.5	1.3
Length of glume c.f. grain	1.8	1.2	1.1	0.9
Intensity of anthocyanin colouration of awn tips	1.6	1.9	1.4	1.0
Ear shape	1.6	1.5	1.2	1.1
Frequency of plants with re-curved leaves	1.6	1.2	1.3	1.0
Spiculation of awn margins	1.4	0.7	0.5	1.5
Time of ear emergence	1.4	1.6	1.7	1.1
Shape of the tip of the sterile spikelet	1.2	1.6	1.1	1.2
Adherence of grain husk	1.0	1.4	1.2	1.0
Average Inflation.	17.0	1.9	1.4	1.1
Coefficient of variation	150%	34%	23%	20%

Considering the example traits, Figure 8-25 shows the QQ plot for the low inflation traits. The presence of significant markers for the anthocyanin trait is still evident, no other trait exhibits significant association against the Bonferroni threshold.

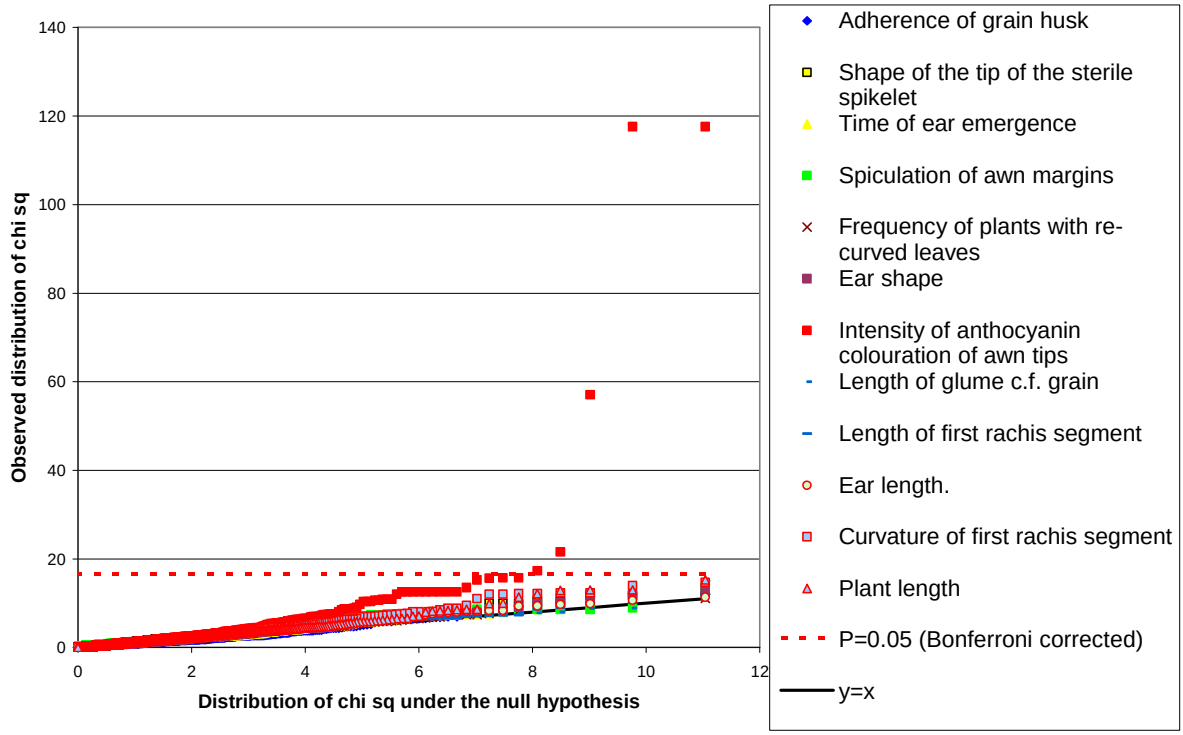


Figure 8-25. Q-Q plot: traits showing low inflation under the simple regression model. Using the MM to estimate association results in a test statistic which follows the null for most of its range for most traits. The anthocyanin competence trait retains its significant markers but no other trait exhibits association significant at the Bonferroni corrected P= 0.05 threshold.

Figure 8-26 shows the QQ plots for the high inflation traits; in this case all traits exhibit association with at least one marker. This contrasts with the previous models where plant growth habit had consistently failed to show association. It could be argued that not using GC as an additional treatment has inflated the significance of these results but reference to Table 8-5 shows that residual inflation for growth habit is low. If GC is applied to these traits the significant markers identified do not change.

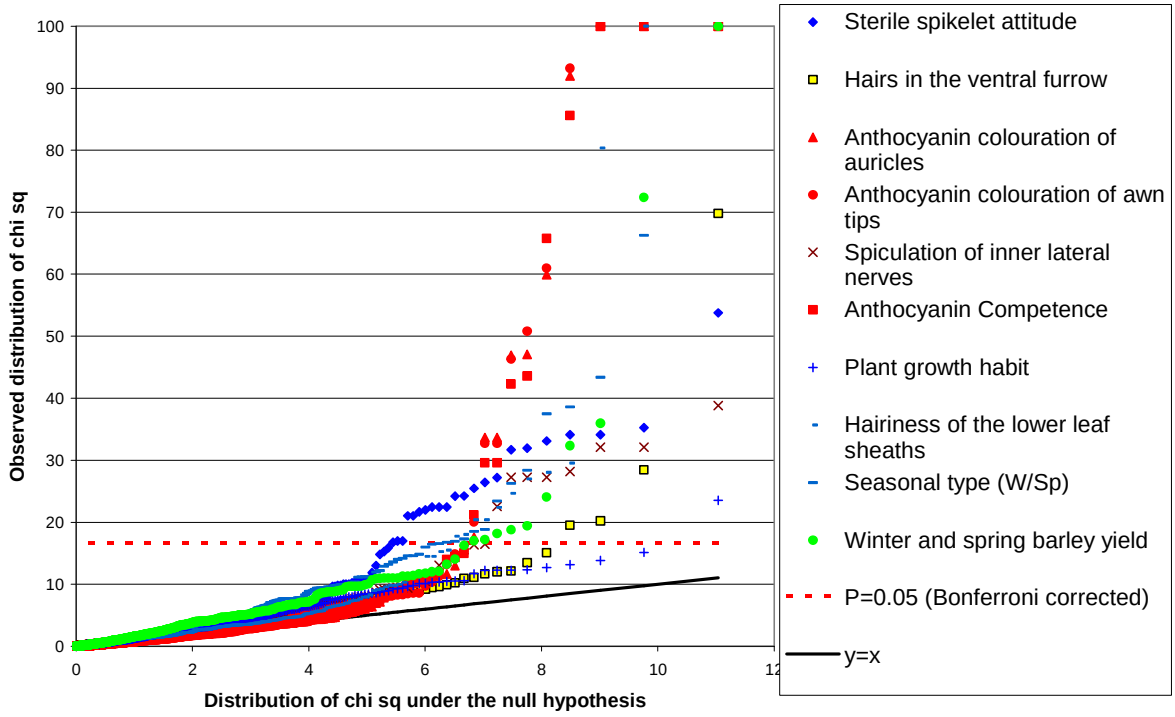


Figure 8-26. Q-Q plot: traits showing extreme inflation under the simple regression model. Note the y axis is truncated at 100 to reveal detail. Using the MM to estimate association, all of these traits exhibit significant association with at least on e marker at the Bonferroni adjusted P=0.05 threshold.

The plots of –log(p-value) over genetic map (Figure 8-27 to Figure 8-29) show very similar features to those exhibited by PC+GC. All association plots are presented in Appendix 10.

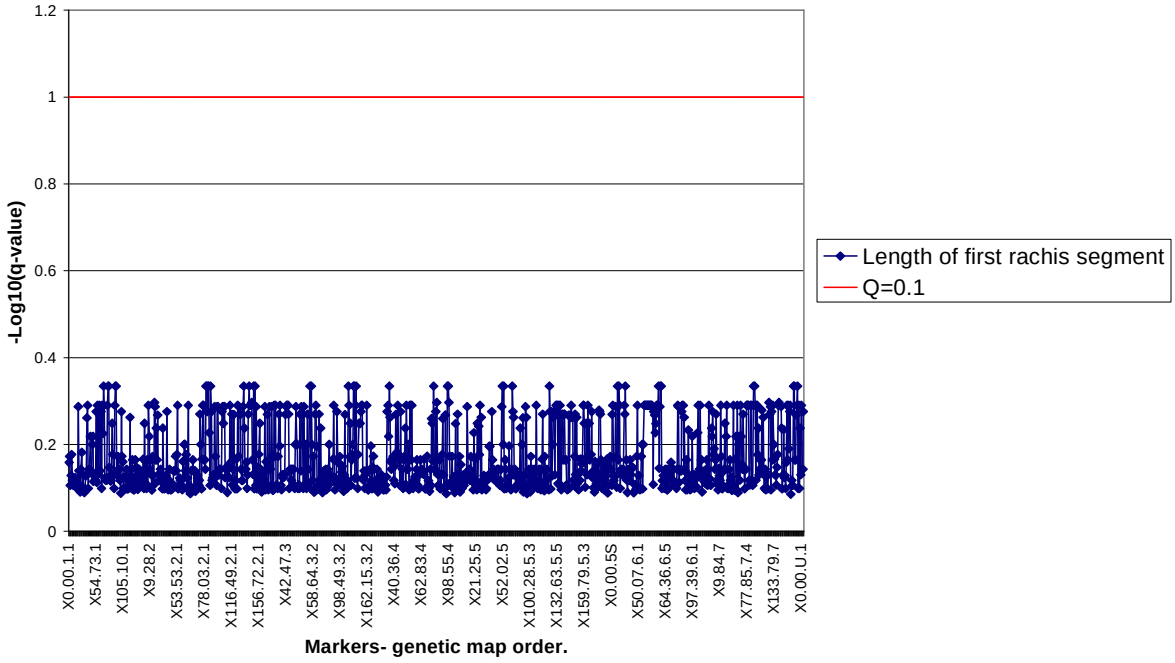


Figure 8-27. Length of the first rachis segment. Q-value over marker panel, markers are in genetic map order. There is no evidence for association with this trait.

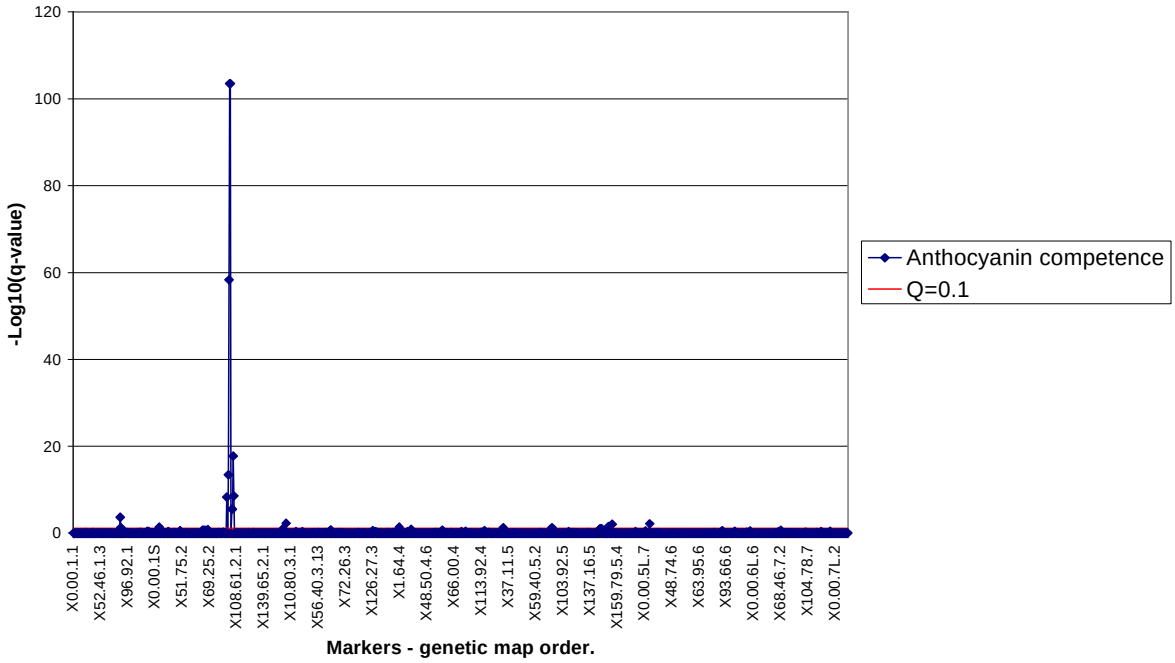


Figure 8-28. Anthocyanin competence. Q-value over marker panel, markers are in genetic map order. The single peak of association seen in simple regression analysis persists; the generalised genome-wide association has been largely controlled.

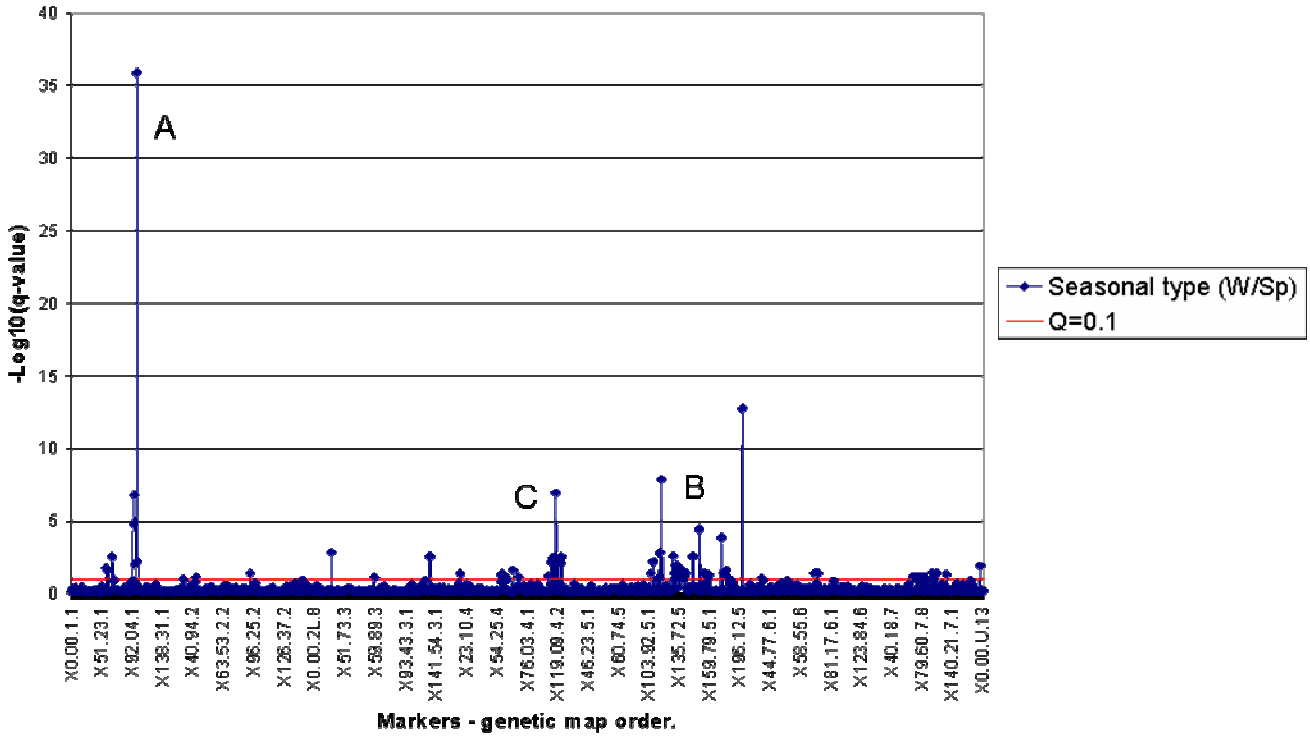


Figure 8-29. Seasonal type. Q-value over marker panel, markers are in genetic map order. The genome-wide association seen in simple regression has been removed; three regions appear to contain significant associations; (A) on chromosome 1 (96cM) the peak is within 10cm of a known dormancy QTL; (B) on chromosome 5 (100-150cM) which straddles VRN-1, a gene responsible for vernalisation response; the physiological mechanism behind 'seasonal type'. (B) is also within a cluster of QTL linked to frost tolerance and (C) on chromosome 4 100-123cM which is within 15 cM of VRN-2, another vernalisation response gene

8.8.1 Summary of Significant Associations

Table 8-4 contains a summary of associations discovered using the four different statistical corrections applied. Significance in this table has been determined using q-values; this contrasts with the published version of the mixed model analyses (Cockram *et al.* 2010) in which significance was determined using a Bonferroni corrected P=0.05 threshold, additionally where ≥ 2 peaks were significant within a 4cM window Cockram *et al.* used this as evidence that the association was less likely to be artifactual. The correspondence between markers significant at q-value ≤ 0.1 and those reported by Cockram *et al.* are indicated by annotation in Table 8-4. The stringent threshold used by Cockram *et al.* was an attempt to differentiate objectively between major gene effects which appeared as strong peaks of association and occasional isolated weaker associations with single markers. The nature of the traits was such that it was reasonable to expect major gene effects for a significant proportion.

A threshold of $Q \leq 0.1$, in these data, is generally less stringent than the P=0.05 Bonferroni threshold (although the relationship between the two scales varies with the distribution of p-values observed in each trait), consequently there are associations identified using q-values which were not reported by Cockram *et al.* These previously unreported associations generally relate to the more quantitative traits (such as length, height and curvature) which is consistent with the use of a stringent significance threshold in the initial analysis. Additionally, in the analyses reported here yield has been included, this trait was not reported by Cockram *et al.*

Of the 32 traits examined (counting all anthocyanin traits as one trait) only nine exhibited no marker-trait association in this study. Of the 20 botanical traits (excluding yield traits) nine exhibited marker-trait associations which confirmed the locations of genes previously identified in bi-parental mapping studies (Table 8-6) and in some cases may have given estimates of genetic map position where previous work had only localised a gene to a chromosome arm (see; Colour of the aleurone layer, Hairiness of the lower leaf sheath, Rachilla hair type in Table 8-6) .

Table 8-6 Significantly associated markers which corroborate previously published bi-parental mapping studies (summarised from Cockram *et al.* 2010; supporting material)

Trait	Location of association peak.	Gene identified in bi-parental mapping.	Location of gene.
Anthocyanin production	Ch2: ~96cM	*ANT2*	Ch2: ~93-97cM
Development of the sterile spikelet	Ch2: ~85cM	*VRS1*	Ch2: ~85cM
Sterile spikelet attitude	Ch2: ~88cM	*VRS1*	Ch2: ~85cM
Row number	Ch2: ~85cM	*VRS1*	Ch2: ~85cM
	Ch4: ~26cM	*Int-c*[1]	Ch4: ~20-30cM
Colour of the aleurone layer	Ch4: ~63cM	*Blx1*	Ch4 (unmapped)
Hairiness of the lower leaf sheath	Ch4: ~111cM	*Hsh*	Ch4 (unmapped)
Rachilla hair type	Ch5: ~100cM	*Srh*	Ch5 (unmapped)
Seasonal type	Ch5: ~ 100-161cM	*VRN-H1*	Ch5: ~ 137cM
Spiculation of inner lateral nerves	Ch2: ~ 78cM	*GTH1*	Ch2: ~ 72cM (est)

[1] These associations have been reported an independent barley variety panel and polymorphisms within *int-c* have been correlated with phenotypic variations. *Int-c* is now known to be an ortholog of teosinte branched 1 (*TB1*), a gene which controls the development of lateral branches with inflorescences in maize. (Ramsay et al., 2011)

We now consider the quantitative traits, height (= plant length) and yield. Yield QTL are reported in the GrainGenes Agronomic QTL consensus map (http://wheat.pw.usda.gov/GG2/index.shtml) in which 14 QTL are identified (if coincident QTL are counted only once) with an average span of 13cM. I counted a co-incidence between associated marker and QTL when the marker falls upon or close to a published QTL; a window of 7cM beyond the published extent of the QTL was allowed since LD extends this far in marker to marker association. Thus an average QTL has an effective extent of ~27cM.

Similarly, plant height QTL are reported in the GrainGenes Agronomic QTL consensus map (http://wheat.pw.usda.gov/GG2/index.shtml) in which 18 QTL are identified (if coincident QTL are counted only once) with an average span of 10cM. If one allows a 7cM window outside the published QTL when declaring coincidence this gives an average QTL size of ~ 25cM.

On this basis a Monte Carlo simulation (10,000 permutations) was used to estimate the null distribution of association discoveries if association discoveries were independent of bi-parental mapping discoveries. The results are summarised in Table 8-7 and show that in the case of yield and plant height the co-incidence between association discoveries and published QTL could easily have occurred by chance.

Table 8-7. Results of simulation in which the positions of QTL and associated markers were allocated randomly.

Trait	No. association discoveries	No. association discoveries co-incident with known QTL	Proportion of simulations in which ≥ co-incident discoveries were made.
Winter Barley Yield	6	2	0.93
Spring Barley Yield	1	1	0.24
Winter and Spring Barley Yield	5	2	0.85
Plant Height	3	2	0.5

Considering the phenotypic variance accounted by significantly associated SNPs (Table 8-8), where genes have previously been identified to account for marker-trait association (Table 8-6) the position of the most informative marker is generally consistent with that gene. The exceptions being sterile spikelet attitude and seasonal growth habit. In the case of seasonal growth habit the more informative marker is on chromosome 1 when the known vernalisation loci are on chromosomes 4 and 5.

In the case of seasonal growth habit (see; Figure 8-24).the chromosome 1 locus is proximate to a dormancy QTL. Given that a major division in the population is between winter and spring types it conceivable that the regression model has largely removed out the signal arising around the causative loci. There is some evidence to support this in Figure 8-8, where association peaks exist in the relevant regions of chromosomes 4 and 5, albeit amongst generalised genome-wide inflation, they have been suppressed in Figure 8-24.

The confounding effect of structure may not have been fully controlled for the combined winter and spring yield trait which exhibits a notably informative marker on chromosome 1, accounting for nearly 50% of phenotypic variation. However, this marker is the same one which appears as the major contributor to seasonal growth habit and so the predictive power for yield simply reflects the yield differential between winter and spring types: it could be viewed as an artefact of structure. Certainly the locus offers no useful predictive power within either the spring or winter groups.

A similar argument can be advanced for the association with yield on chromosome 2 ~82-85cM which is coincident with *VRS1* and so is likely to reflect the yield differential between

6-row and 2-row barley. Again, the yield differential is real but the breeding interest would be for markers predictive of yield gain within the seasonal type/row number classes.

Some other yield associations ([redacted]) are proximate to *VRN2* and *VRN1* respectively, however there remains one region (Chromosome C ~ 80-85cM) which is not (1) obviously associated with structure, (2) consistently associated with yield and (30 also co-incident with a published QTL for yield. This, at least, may be an association resulting from LD with the QTL.

There is a trend in Table 8-8 such that highly heritable traits are also those for which significant associations are discovered. Figure 12-10 shows this trend in graphical form. It is notable that the trend is somewhat consistent with the predictions for statistical power made in 6.4 where high power of 90% or more was predicted for Mendelian traits with heritability as low as 50%.

Table 8-8. Heritability and Proportion of phenotypic variance (V_P) accounted by SNPs for significantly associated trait/marker combinations together with heritability of traits with no significant association. (Adapted from Cockram *et al.* 2010)

Trait	Marker position	GWA significance [-log10(p)]	Trait h^2	Trait V_P	SNP V_P	SNP h^2
Sterile spikelet attitude	Ch1: ~55cM	12.65	0.48	0.33	0.08	0.25
Sterile spikelet attitude	Ch2: ~88cM	8.29	0.48	0.33	0.05	0.16
Seasonal growth habit	Ch1: ~96cM	38.83	0.80	0.12	0.08	0.67
Seasonal growth habit	Ch5: ~122cM	10.3	0.80	0.12	0.02	0.20
Sterile spikelet development	Ch2: ~85cM	17.8	0.10	2.18	0.98	0.45
Awn anthocyanin coloration	Ch2: ~94cM	113.47	0.58	9.79	8.18	0.84
Awn anthocyanin intensity	Ch2: ~94cM	26.67	0.35	2.47	0.85	0.34
Auricle anthocyanin coloration	Ch2: ~94cM	103.91	0.57	9.84	8.01	0.81
Auricle anthocyanin intensity	Ch2: ~94cM	29.53	0.45	3.78	1.28	0.34
Lemma nerve anthocyanin intensity	Ch2: ~94cM	37.84	0.59	4.21	2.09	0.49
Grain lateral nerve speculation	Ch4: ~26cM	4.11	0.63	5.24	0.52	0.10
Grain lateral nerve speculation	Ch4: ~54cM	3.52	0.63	5.24	0.19	0.04

[1] Identity of markers obscured because permission to publish has not been given.

Grain lateral nerve speculation	Ch2: ~78cM	7.83	0.63	5.24	1.05	0.20
Grain lateral nerve speculation	Ch2: ~151cM	4.30	0.63	5.24	0.22	0.04
Grain lateral nerve speculation	Ch5: ~180cM	4.28	0.63	5.24	0.09	0.02
Ear row-number	Ch1: ~75cM	8.2	0.8	0.04	0.00	0.14
Ear row-number	Ch2: ~85cM	7.1	0.8	0.04	0.00	0.12
Ear row-number	Ch4: ~26cM	68.87	0.8	0.04	0.03	0.78
Ear row-number	Ch4: ~117cM	6.26	0.8	0.04	0.00	0.02
Ear row-number	Ch6: ~105cM	4.8	0.8	0.04	0.00	0.12
Aleurone color	Ch4: ~63cM	67.38	0.72	0.50	0.37	0.74
Hairiness of leaf sheath	Ch4: ~111cM	39.9	0.61	5.17	2.52	0.49
Grain rachilla hair type	Ch5: ~100cM	13.22	0.58	0.14	0.03	0.20
Ear attitude	Ch5: ~135cM	5.49	0.23	2.19	0.21	0.10
Grain ventral furrow hair	Ch4: ~26cM	3.98	0.55	8.32	0.93	0.11
Grain ventral furrow hair	Ch6: ~1cM	7.02	0.55	8.32	0.90	0.11
Plant length	Ch2: ~87cM	3.48	0.12	1.67	0.07	0.04
Plant length	Ch4: ~51cM	3.50	0.12	1.67	0.09	0.05
Plant length	Ch4: ~81cM	4.05	0.12	1.67	0.09	0.05
Spring barley yield	ChC: ~82cM	3.56	0.39	0.01	0.00	0.10
Winter barley yield	ChC: ~85cM	4.70	0.4	0.11	0.02	0.14
Winter barley yield	ChE: ~56cM	4.75	0.4	0.11	0.01	0.13
Winter barley yield	ChG: ~26cM	4.68	0.4	0.11	0.03	0.16
Winter barley yield	ChG: ~55cM	3.06	0.4	0.11	0.01	0.09
Winter barley yield	ChF: ~117cM	3.10	0.4	0.11	0.00	0.04
Winter and spring barley yield	ChA: ~96cM	26.0	0.66	0.07	0.03	0.47
Winter and spring barley yield	ChC: 50-82cM	3.20	0.66	0.07	0.00	0.03
Winter and spring barley yield	ChG: ~117cM	16.75	0.66	0.07	0.02	0.23
Winter and spring barley yield	ChF: 150-170cM	4.47	0.66	0.07	0.01	0.13
Winter and spring barley yield	ChD: ~105cM	3.75	0.66	0.07	0.00	0.04
Adherance of the grain husk	-	-	0.00	-	-	-
Shape of the tip of the sterile spikelet	-	-	0.07	-	-	-
Time of ear emergence	-	-	0.27	-	-	-
Spiculation of awn margins	-	-	0.00	-	-	-
Frequency of plants with re-curved leaves	-	-	0.14	-	-	-
Ear shape	-	-	0.15	-	-	-
Length of glume	-	-		-	-	-

c.f. grain			0.14			
Length of first rachis segment	-	-	0.34	-	-	-
Ear length.	-	-	0.16	-	-	-
Curvature of first rachis segment	-	-	0.22	-	-	-
Collar type	-	-	0.09	-	-	-
Ear density	-	-	0.19	-	-	-
Disposition of lodicules	-	-	0.41	-	-	-

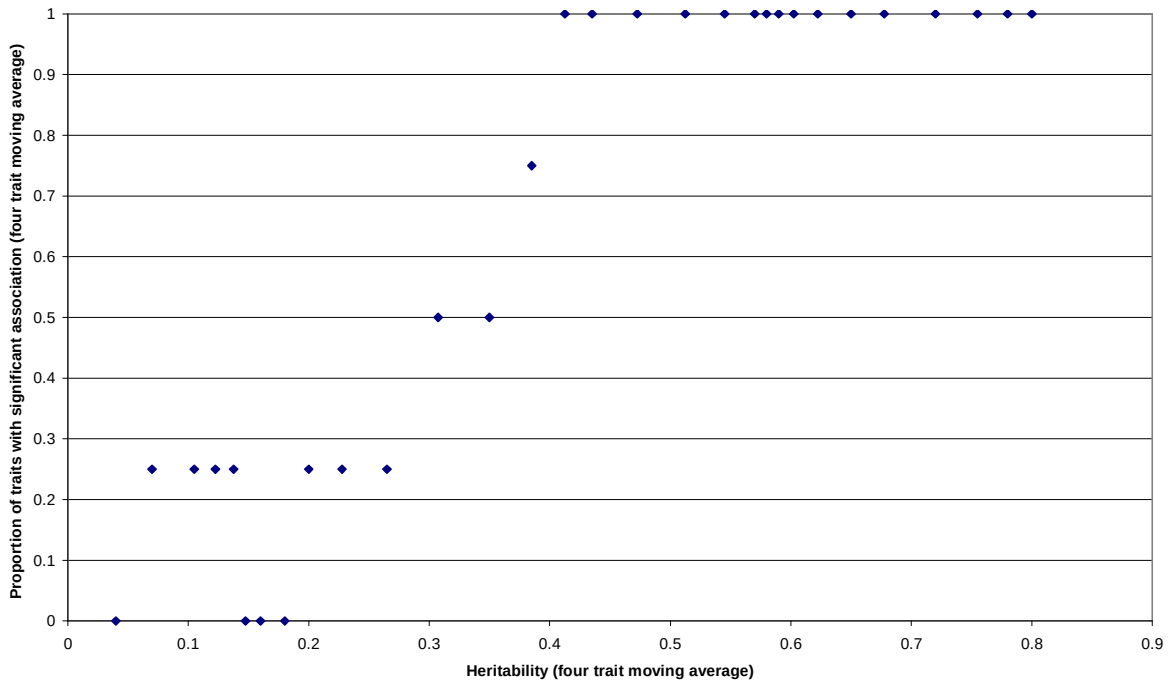


Figure 8-30. Moving average of proportion of traits exhibiting significant association over ascending heritability of traits.

8.9 Confirmation of association mapping discovery and identification of a functional polymorphism for ANT2.

Cockram *et al.* (2010) report the exploitation of a bi-parental mapping population and the exploitation of Rice-Barley synteny to identify candidate genes for anthocyanin competence on chromosome 2 of barley proximate to the association peak at 96cM. This led to the identification of a potential causative polymorphism within *ANT2*. In this work I contributed further association mapping and analysed the profile of LD over the genetic map of chromosome 2. Bioinformatics and additional molecular biology was conducted by James Cockram.

The peak of significant association with anthocyanin competence on chromosome 2 was confirmed to be co-incident with *ANT2* by bi-parental mapping. A doubled haploid mapping population derived from a cross between the varieties Saffron (anthocyanin incompetent) and Retriever (anthocyanin competent) was used to show that the trait was Mendelian and mapped to chromosome 2 and co-segregated with two markers (11_21175 and 11_21007) which mapped to the same genetic interval identified by association mapping of the trait. This served to show that the association discovery was not an artefact of structure or kinship since these confounding factors are absent in bi-parental mapping. The next step was to identify genes in the region identified which might control the anthocyanin trait.

The barley genome is currently not sequenced. This was a potential constraint on any attempt to use the association discoveries to identify genes or causative polymorphisms; this because the reliance on recombination renders fine scale genetic mapping in slow-growing plants impractical due to cost. A physical sequence, in contrast, gives unambiguous ordering of genes and re-sequencing can reveal SNPs which may be exploited to generate haplotype data in association mapping panels, the haplotypes being anchored in the known physical map.

The lack of sequence data in this crop was largely overcome by exploiting synteny with rice. Figure 8-31 shows that the position of the most strongly associated marker for anthocyanin competence on barley chromosome 2 could be mapped to a homologous region of rice chromosome 4. The sequenced rice genome could then be used as a basis for identifying candidate genes.

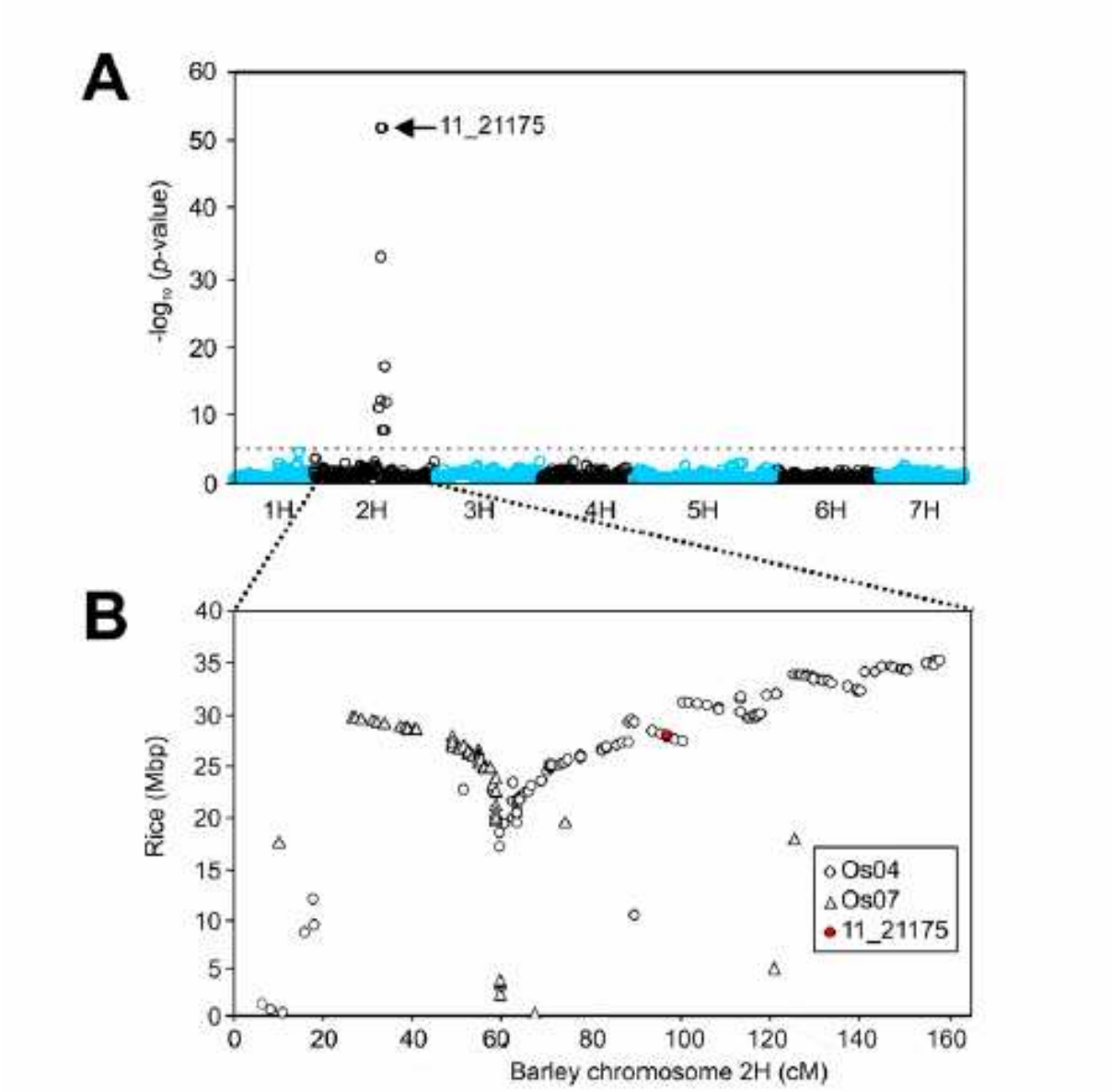


Figure 8-31. Exploiting synteny between barley and rice. (A) Association mapping showed a clear peak of association with anthocyanin competence in chromosome 2 of barley, the most strongly associated marker is identified. (B) Sequence homology between the barley SNPs and the sequenced genome of rice shows barley chromosome 2 could be said to comprise segments of rice chromosomes 2 (○) and 7(Δ) and that the order of the SNPs in the physical map of rice is consistent, over short distances, with the corresponding genetic map order in barley. However, there is compelling evidence for inversions between the two species. The centromere is at ~60cM and the distortion between physical and genetic maps at this point reflects the rarity of recombination events in the centromeric region; essentially the genetic distance scale is shortened relative to physical distance at this point. (Figure reproduced from Cockram *et al.* 2010)

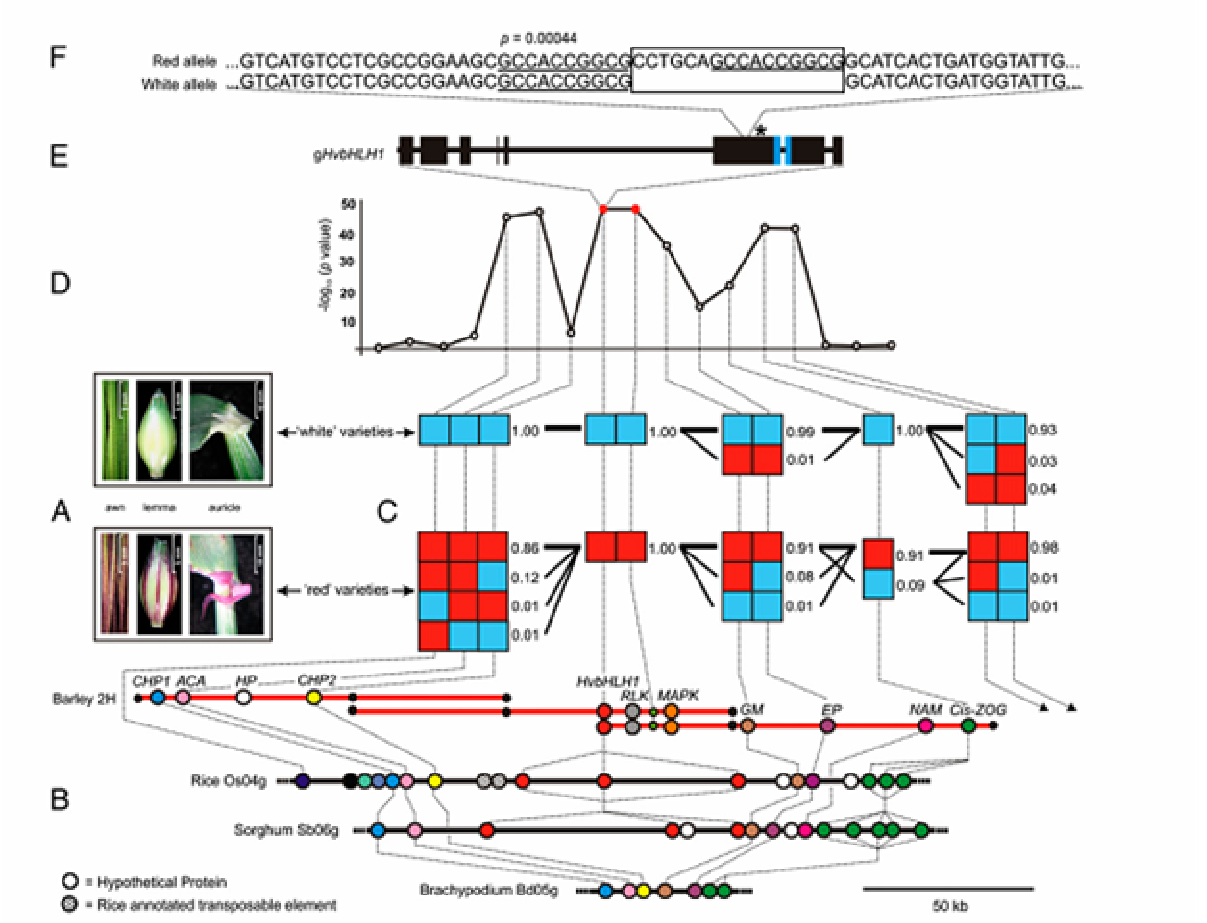


Figure 8-32. (A) Illustration of anthocyanin incompetence (upper) and competence (lower). (B) Alignment of genes in sequenced genomes of Rice, Sorghum and Brachypodium compared to the sequenced segment of barley genome containing the major anthocyanin association. (C) Haplotypes observed in additional genotyping within the sequenced segment. (D) Significance of association over the sequenced region. (E) The structure of *HvbHLH1* showing the site of the 16bp deletion believed to be the cause of anthocyanin incompetence in barley. (Reproduced from Cockram *et al.* 2010.)

Bacterial artificial chromosomes (BACs) of barley spanning the region of interest were sequenced and aligned with the corresponding region in rice. This served to confirm that region comprised barley homologues of the rice genes, established their order, identified any duplications of genes and any inserted DNA which might have contained additional genetically active sequence. (B) shows the alignment of genes on the rice physical map compared to the corresponding genes in the, now sequenced, section of the barley genome. The order of genes is largely conserved between the two species, additional confirmation was subsequently obtained by alignment with Sorghum and Brachypodium sequences.

Having established that there were several genes within the candidate region additional genotyping of 90 barley cultivars revealed SNP and InDel polymorphisms which were developed as markers and used to genotype the entire panel. This allowed finer scale association mapping over the sequenced region. (D) shows the statistical significance of association for these additional markers and (C) shows a graphical representation of the observed haplotypes. It was noted that the two markers identified by red dots in (D) exhibited polymorphisms which were unambiguous predictors of anthocyanin competence in this panel. One of these markers originated in the gene *HvbHLH1*.

Considering the putative function of the genes in the region identified *HvbHLH1* as a candidate for *ANT2* since it encodes a basic helix-loop-helix (*bHLH*) DNA binding protein which is a previously reported (Sweeney *et al.* 2006) feature of transcription factors regulating pigment synthesis.

The protein sequence of the candidate barley *bHLH* was compared with *bHLH* from the anthocyanin pigmentation pathways of petunia, antirrhinum, maize and arabidopsis showed the barley protein most related to its maize homolog. RT-PCR was used to show that *HvbHLH1* was expressed in both Saffron and Retriever.

Sequencing *HvbHLH1*in 90 varieties revealed 69 polymorphisms forming four haplotypes in which haplotype 1 was exclusive to anthocyanin incompetent varieties. Genotyping the non-synonymous polymorphisms in *HvbHLH1* identified a 16 bp deletion in the genetic sequence resulting in a premature truncation of the protein. Over the entire panel this polymorphism proved a perfect predictor of anthocyanin status.

Thus the *ANT2* gene was identified as *HvbHLH1*, a transcription factor, which is universally expressed but exists in an inactive (truncated) form in anthocyanin incompetent varieties. Essentially the anthocyanin biosynthesis pathway is switched off by the absence of active *HvbHLH1*.

Cockram *et al*. present some evidence that the mutation is relatively recent since it is not seen in wild accessions but is seen in land-race varieties (strains of barley selected and developed by farmers; pre-dating modern breeding). Additionally the pattern of LD over genetic map is different in wild-type (competent) varieties than incompetent varieties. Specifically LD is higher in the incompetent group supporting the hypothesis that there is a recent common ancestor for the incompetent group (Figure 8-33).

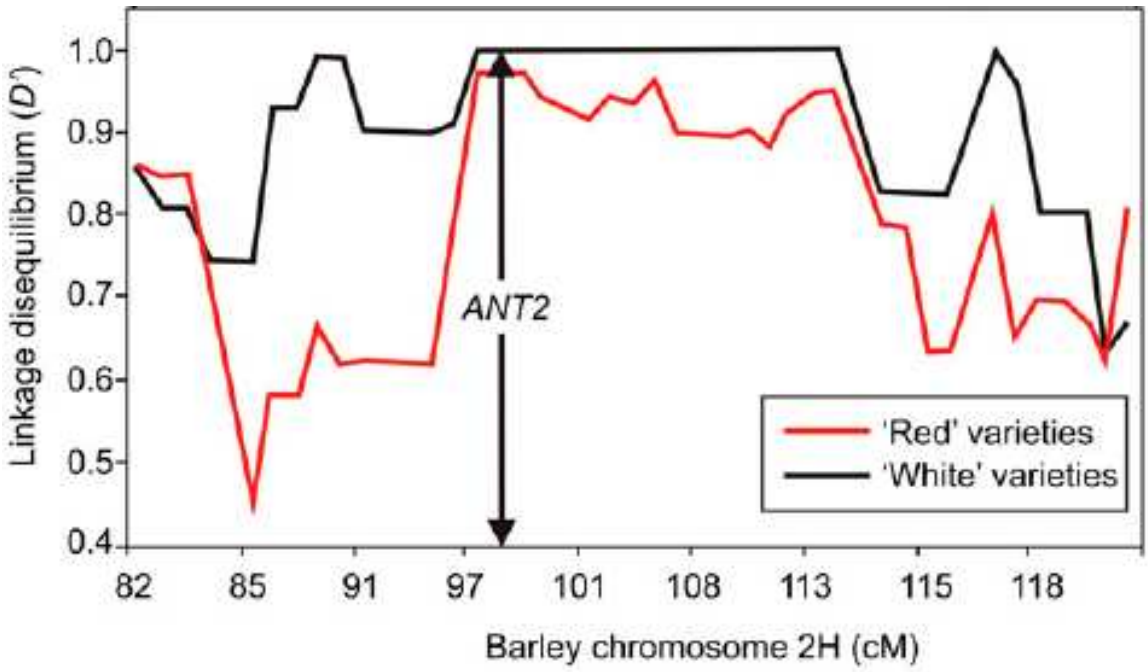


Figure 8-33. The pattern of between adjacent marker LD (D') over genetic map position for both anthocyanin competent (red line) and incompetent (black line) varieties. Moving average over 4 markers.

8.10 Conclusion.

There is evidence for genome-wide inflation of the association test statistic for many of the tested traits. This is at least partially accounted for by models of structure or kinship in which the inflation (measured as inflation of the median of the test statistic for a trait) increases with the amount of phenotypic variation which can be accounted for by the structure model. This relationship appears most consistent in the case of kinship models where low inflation and low predictive power consistently occur together. This suggests that one reason why structure models gave an excess of false discoveries in power simulation compared to kinship models.

For highly heritable Mendelian traits there is a general consistency in terms of the associations discovered over all regression methods. In many cases these discoveries map to locations already known to contain relevant genes or QTL. In some cases the association mapping appears to have improved resolution considerably over previous mapping (see in particular *Blx1, Hsh* and *Srh* in Table 8-6).

The best control of inflation of the test statistic was evident using the mixed model and since power simulations had already shown this to be the most efficient method, in terms of the ratio of the number of discoveries made to those expected under the null, this method was used for association mapping.

Although previously mapped genes had been shown to co-map with the association discoveries for a number of Mendelian traits, this cannot be viewed as unequivocal evidence that association mapping for these traits in this panel had been successful. In fact, for two quantitative traits, despite co-incidence of published QTL and associated markers, the success rate of association is indistinguishable from that under the null which is exactly what was predicted in the power simulations for quantitative traits with heritability of 0.5.

Additional to this thesis one of the association discoveries reported has been further examined and a causative polymorphism for anthocyanin competence has been proposed. This should certainly encourage further work on others of the Mendelian traits; colour of the aleurone layer, hairiness of the lower leaf sheath and rachilla hair type were all only crudely mapped before this work but it may now be possible to exploit co-linearity with related, sequenced, genomes to at least identify candidate genes.

As an illustration of the use of association mapping as a tool to signpost sites for further investigation this barley analysis has been very successful. The analysis has also served to confirm the limitation of statistical power in the panel since there was a general failure to detect confirmable associations in quantitative traits – even when heritability was relatively high.

9 Discussion

In this thesis I report association studies conducted in wheat and barley exploiting historical phenotype data and germplasm from seed-banks.

In wheat I have examined four quantitative traits and found statistically significant associations with markers for three of them. In barley I found associations with 27 out of 37 traits, 34 of the traits were botanical characters known to give a phenotype largely independent of environment.

9.1 The experiments.

The association mapping experiment in wheat was built up from first principles, including sample acquisition, DNA extraction and analysis of historical phenotype data. Notably there was no single genetic map for DArT markers in wheat and the analyses I performed to consensualise the multiple map fragments provided by Triticarte (the genotyping sub-contractor) appears to have been successful in that my map agrees broadly with that of Crossa (2007) published subsequently; I have presented some evidence that my map is superior in its ability to predict D' in my panel.

Based on power simulations I considered MM analysis to be the best analytical approach and the results of association analysis in both wheat and barley panels support this. In wheat the MM results show almost complete control of genome-wide inflation thus avoiding the need for GC which saps statistical power. Pragmatically, the MM discovers more associations (Table 7-15 to Table 7-18) than other methods and those associations include ones coincident with published QTL. Taking the MM discoveries for protein, yield and HFN, these appear to be distributed in the genetic map in a non-random manner with respect to previously published QTL – that is they are coincident more often than one would expect by chance alone. In addition commercial plant breeders have commented that the locations of the associations with HFN which I report here are not only coincident with published QTL but some are also coincident with QTL from their own unpublished research. I have provided map and marker information to third parties for further evaluation.

In barley the panel was provided pre-formatted by SCRI with phenotypes of botanical characters summarised by colleagues. These data provided a useful contrast to the wheat panel with simpler traits and more extensive panels offering much greater statistical power

both in simulation and in fact. Again, the MM proved superior in terms of control of inflation and the number of associations found (Table 8-4). The association with anthocyanin competence was used to demonstrate a profitable alliance between statistical genetics, bioinformatics and molecular genetics allowing the research team to move from a statistical association in the un-sequenced barley genome into homologous regions of the rice genome. We then used the rice sequence as a template for sequencing and genotyping barley through a number of candidate genes before eventually showing that the polymorphism which perfectly predicted of anthocyanin status was a functional variant of *HvbHLH1,* a transcription factor.

In barley there were at least five other traits showing similarly strong single peaks of association which could be similarly investigated. Indeed one of these, *Int-c,* was mapped with increased precision in this association study (see Table 8-6) and this new location has subsequently been confirmed as part of the full characterisation of the gene by Ramsey *et al.* 2011.

9.2 Comparison with human studies.

It is impossible to discuss these results without making some comparison with association mapping studies in human populations since, for the general reader, this will be the point of reference. It would be fair to say that panels and results reported here contrast markedly with genome-wide association studies in human populations both in terms of panel size and the proportion of phenotypic variance accounted by SNPs. This difference is probably attributable to several factors:

In human association studies the phenotype is generally a disease state recorded on a case/control basis. However, for some diseases the diagnosis may be ambiguous and for the majority, disease state will be influenced both by environmental factors and by genotype x environment interactions. These sources of error will contribute to the environmental component of the phenotypic variance and tend to reduce the heritability of traits within the experiment. The deflation of trait heritability may be counteracted by larger panel sizes which increase the precision with which effect sizes can be measured.

Even though disease state may be recorded in binary format there are often many loci influencing the trait, each with small effect. Thus the individual SNP heritability may be low and the power to detect any one such SNP will also tend to be low. That SNPs are successfully and verifiably identified as being associated with disease reflects, to some extent, the variance of the SNP effect; in some panels some SNPs will exhibit more than their

average effect and may become statistically significant as a result – this is the ‘winner’s curse’. Additionally, large panel sizes allow small effects to be detectable over the base noise.

In contrast, association mapping in agricultural crops potentially benefits from phenotype data which have been collected in randomised replicated trials grown at multiple locations over several years. The consequence being that the environmental variance has been partitioned in to year effect, site effect, block effects, etc. Additionally, replication of genotypes is explicit from experimental design. This replication serves to reduce the environmental variance component of phenotypic variance.

The recent genetic history of agricultural crops has an influence on association mapping. For example, there are many quantitative traits in crops (for example; yield) which will be influenced by many loci. In wild species these multiple loci may well behave as the multiple loci influencing human disease risk – each having a small effect. In the development of agricultural crops, however, there will be a tendency to select for some QTL which have large effects on traits and to fix the beneficial variant in the breeding population. The QTL which have been fixed in this way are in some cases beneficial because they complement the local environment. For example one strategy identified to increase yield of wheat in drought environments (Jordan *et al*. 1983) is to select for short growing season, this is, counter-intuitively, beneficial for yield since it allows the crop to reach maturity before the annual drought kills the plants, this is the core strategy for the rain-fed wheat crop in S. Australia; in contrast in N. Europe water is not limiting until mid-summer and highest yields are obtained from long growing seasons with an entire pool dedicated to producing material suitable for autumn sowing; that is to be frost hardy and to suppress flowering for the winter months. Other beneficial QTL may have smaller effects and have become fixed as a consequence of selection and drift without their location or effect ever having been noted. Other QTL influencing yield may never have been fixed in breeding populations and these are available for discovery by mapping techniques. It is worth noting that founder effects, selection and drift in developing populations may have resulted in the loss of potentially beneficial QTL which may still exist in wild populations. The fixation of a subset of potentially beneficial QTL can be viewed as a manifestation of both drift and the ‘winner’s curse’ in as much as the traits fixed by selection may simply have been those whose effect was large enough to be noticed in the conditions under which selection took place. QTL fixed as the crop was domesticated, selected and bred (or fixed in the experimental panel) cannot, of course, be detected using association mapping.

So, unlike human populations, there may be no fundamental 'truth' about what constitutes a beneficial allele, it will depend where the crop is grown. In addition there are probably many beneficial alleles which have become lost to breeders, essentially by accident. The implication is that panels for association mapping of traits with significant environmental interaction or adaptive significance should, as a first principle, be conducted in panels which are matched for the adaptation. This will restrict both the size of variety panels which can be assembled from gene-banks and, the universal applicability of any associations found.

In the barley association studies presented, the majority of the traits were those used to define varieties using botanical appearance, consequently these were characters which were unambiguously defined and expressed in a reproducible manner irrespective of environment; they would tend to be highly heritable even in very small trial programs, and indeed variety descriptions may be established over as few as two years with one trial in each year. Of those botanical traits which are not Mendelian there is a parallel with the human disease case in that the phenotype is simply stated but there may be many loci influencing the trait. The Mendelian characters, in contrast, should exhibit a single locus accounting for a high proportion of phenotypic variation.

In addition to the accuracy with which the genetic variation may be determined, the extent of LD in panels of agricultural crop varieties is greater than that found in outbred natural populations, such as humans. The extent of LD in the winter wheat panel is of the order of 5-10cM and in the combined spring and winter barley panel, 10-20cM. This contrasts with the human situation where LD is typically fully decayed at 200 kbp (for example; Dawson *et al.* 2002) which equates to ~0.2cM. The larger extent of LD will mean that associations may be detected over greater genetic distance than in the human case and so, in principle, smaller marker panels may be used. The extent of LD and the distribution of between marker-pair distances presented in 5.1.2 and 5.2.3 show that the marker panels used here sample the expected extent of LD and should have some capacity to detect association between markers and QTL.

It is, therefore, less surprising than initial impressions would suggest that these small panels should have uncovered some associations and that the associations in some cases are very strong (the Mendelian botanical traits). In barley there was some evidence that a trait imperfectly correlated with seasonal growth habit (dormancy) was associated with yield but this probably resulted from the imperfect secondary correlation with seasonal growth habit

which itself correlates with a large yield differential (the difference between winter and spring sowing).

9.3 Methods, power and false discovery.

When the research presented in this thesis commenced there was no clear consensus as to the best method for association mapping in crops. Regression techniques such as SA and PC correction were extant and the MM approach had been demonstrated in maize and *Arabidopsis*. The literature contains many examples of experiments which appear to have been conducted opportunistically on existing data, the regression methods used for analysis are often stated with no explicit comment as to why one was chosen over another. Notable exceptions to this were Zhao *et al.* 2007 and Yu *et al.* 2006 who estimated power over a range of available methods in their panels. These authors pointed towards mixed effects modelling as being superior although it was not clear if this would be true in naturally inbred species.

In this thesis I used power as metric for differentiation and evaluation of methods and concluded that power without a parallel estimate of false discovery rate was of somewhat limited value; reducing the argument to absurdity one might simply assert that all markers are associated with a trait – this would give 100% power but it would also give a list of discoveries which were almost all false. Conversely, minimising false discovery erodes power and does not guarantee that remaining true discoveries are due to linkage – the alternate hypothesis in association mapping. This relationship is often summarised as a receiver operating characteristic (ROC) curve in which the true positive rate is plotted over the false positive rate.

Estimating empirical false discovery in an association panel is not possible since even if the phenotype is spiked with additional genetic effect it is only possible to test whether the source(s) of that effect can be found by association; other discoveries made at the same time may or may not be false. The q-value gives a theoretical solution to the problem by estimating the proportion of p-values under the null and alternate hypotheses at a given significance threshold. However, although the approach quantifies p-values under the null it is not able to offer any guidance as to which of a number of alternate hypotheses are in play. Thus a low q-value for a marker might just as easily be due to inflation due to (for example) uncontrolled structure effects as to genuine association.

I used the fully simulated phenotypes in wheat, barley and the simulated cereal population to attempt to calculate the empirical FDR for a given method under various circumstances. The advantage of the approach was that if a rule was stated – that a discovery within a certain genetic distance of the known causative polymorphism was true – then other discoveries could be defined as false. Even this is not perfect since it is conceivable (although unlikely) that associations could arise in close proximity to the causative polymorphism for reasons other than linkage and would be incorrectly counted as true, and *vice versa*.

With a fully defined genetic architecture for a trait it is possible to calculate the efficiency of competing analytical approaches and the ratio of power from the method which makes *n* discoveries to the power from a null model in which the position of *n* discoveries are randomly assigned in the same genetic map (the power ratio). This ratio proved useful since it measures gain over the null and this is the core objective of any regression analysis.

In these data the best choice of regression approach was not entirely obvious even with an efficiency score and a power ratio although it was clear that the MM was consistently best or among the best. Since the choice of method was not unambiguous, I performed association analyses with a range of techniques.

A useful piece of further work would be to extend the power simulations to include other configurations of the MM such as those in Yu *et al.* 2006 where additional covariates such as the **Q** matrix from STRUCTURE are included. Zhao *et al.* 2007 also report an approach with **K** truncated to have negative values set to zero and **Q** as a covariate. These approaches seem unnecessary since **K** without truncation includes information from both structure and kinship. Yu *et al.* report a modest gain in power when **Q** is added but it may be that this is offset by a rise in FDR. Bringing these alternative models together and comparing efficiency and power ratio as I have with other methods may reveal ways to further improve the analysis.

In the analysis of the barley panel 8.4.2, I was able to consider the relationship between inflation of the test statistic in regression with the ability of structure/kinship models to predict phenotype. The result was interesting in that structure models predicted phenotypic variation well in circumstances where the inflation was low. I interpreted this to be due to groups in the panel defined by the presence of local similarity in the genome rather than a genome wide effect. It seems likely that applying a structure model defined in this way in a genome-wide correction will understate the residual phenotypic variation against which trait-marker association is to be measured. The kinship model, in contrast to structure, was unable

to predict phenotype when the inflation was low suggesting that the genome-wide average similarity which defines the genetic correlation between two individuals is not so influenced by local similarities. This was consistent with the ability of STRUCTURE to find sub-populations in randomly mated individuals 3.3.3 and the disappearance of structure when the chromosome containing the 1B1R introgression from rye was excluded from the wheat marker panel. There are thus some hints as to why the mixed model, using kinship as a random effect, may offer more consistency in performance over phenotypes.

Ultimately, however, a regression method must be used in data where the full architecture of the trait is not known. In this case power is an important consideration since it allows the experimenter to predict in advance how successful an association mapping experiment is likely to be. This process can inform experimental design while an estimate of power made in parallel to an association study certainly provides useful information the interpretation of association results. For example if many associations are discovered in an under-powered experiment it suggests the wrong alternate hypothesis is in play; if no associations are made in a moderately powered experiment it might suggest individual QTL effect sizes are very low. This use of power as a quality assurance step remains uncommon in plant association studies although this is changing (Zhao *et al.* 2007, Yu *et al.* 2006 and Cockram *et al.* 2010).

In this study power was estimated both in the simulated phenotypes and by perturbing existing phenotypes (Chapter 6), the results are summarised in Table 9-1.

Table 9-1 Power to detect association (using MM) in the available panels; simulated phenotype (Q≤0.1).

Species	Proportion of variation explained by a QTL	Proportion of Loci detected (association noted within 10cM)	Power ratio (gain over null discovery rate)
Barley	0.9	0.9	14
	0.5	0.85	10
	0.1	0.30	4.5
Wheat	0.9	0.55	9.5
	0.5	0.45	8
	0.1	0.2	2.5

Table 9-1 shows there is ample power in the barley experiment to detect Mendelian traits when the marker effect is high. The experimental outcome (summarised in Figure 9-1) is entirely consistent with this prediction.

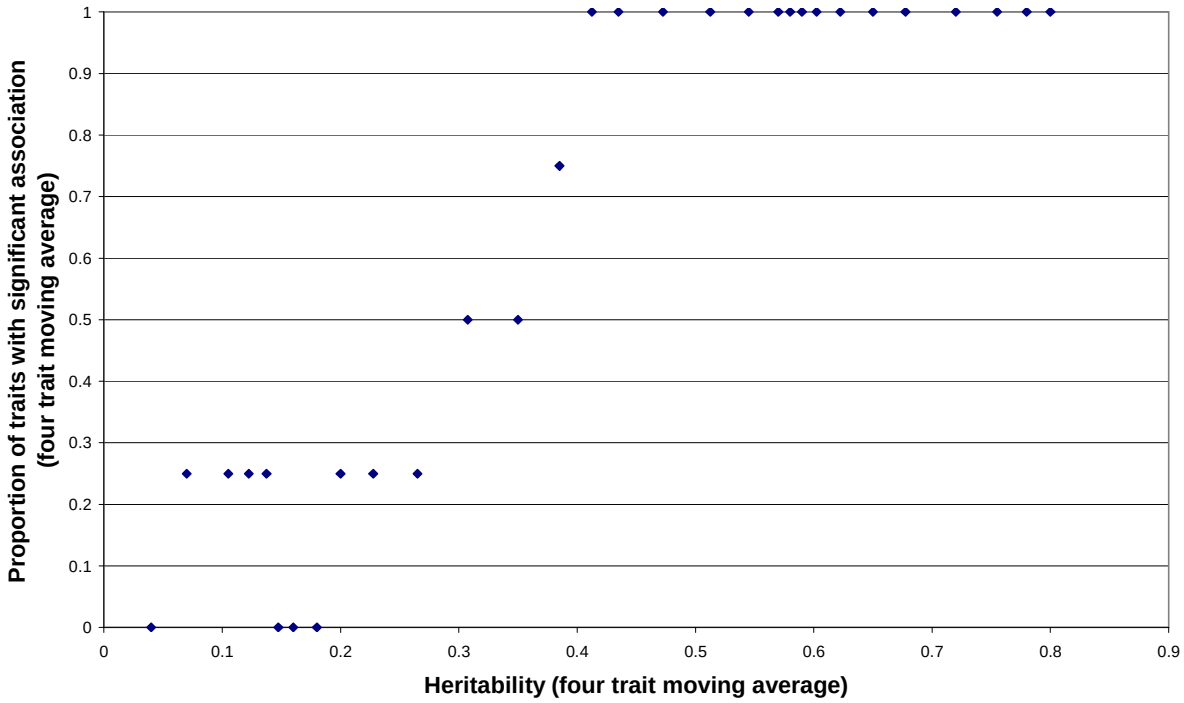


Figure 9-1. Moving average of proportion of traits exhibiting significant association over ascending trait heritability.

9.4 Quantitative traits.

Turning to the power to detect perturbation to the available quantitative traits for wheat and barley (Table 9-2). I found the power to detect associations in wheat is broadly consistent with the prediction from the simulated phenotype (taking 10% of variation explained by a QTL as being a quantitative trait). However the power for the barley yield traits is considerably higher than expected from simulation.

Table 9-2 Power to detect association (using MM) in the available panels; perturbed phenotype (Q threshold as indicated).

Crop and trait	Power (Q≤0.1)	Power (Q≤0.3)
Wheat yield	0.2	0.3
Wheat HFN	0.3	0.43
Wheat protein	0.2	0.42
Wheat Sp.Wt.	0.12	0.28
Spring Barley yield	0.5	-
Winter barley yield	0.65	-
Winter + Spring yield	0.85	-

Yield is a classic example of a quantitative trait and one would expect it to be controlled by multiple QTL of small effect. However, mixing together populations (for example the spring and winter barley set) can result in distinctly bi-modal distributions for the trait and there can be very strong genetic correlations generated by the mixing – for example the Mendelian trait of vernalisation requirement. However, it could be argued that knowing there is a yield differential between winter and spring barley is of no value since a variety is, to all practical purposes either in one group or the other – it is differentiation within the group that is important. Also bulking together pools adapted to different environments is likely to highlight adaptation differences rather than useful QTL. Accepting this argument, it may be better to conduct an association study within a sub-population rather than over multiple sub-populations. There is a counter argument to this; since QTL influencing yield may have become fixed independently in winter and spring pools and since these pools share environment for at least some of the growing season, a high yielding spring barley may be exploiting a different set of beneficial QTL than a high yielding winter barley. If this is the case then an association study spanning both populations when compared to association studies within sub-populations might uncover unfixed QTL common to winter and spring and also QTL fixed in one or other group. It is conceivable that some beneficial yield QTL in spring barley could offer advantages to the winter crop and *vice versa.* In this respect, the

marker noted on Chromosome G at 26cM is significantly associated with winter barley yield but is fixed in the spring barley panel. It is also notable that the spring barley panel is fixed for the allele associated with lower yield in the winter crop. No other significant marker is fixed in one panel but not in the other.

Table 9-3. Positions, statistical significance and heritabilities of markers associated with yield in barley.

Trait	Marker position	GWA significance [-log10(p)]	Trait h2	Trait VP	SNP VP	SNP VP / Trait Vp
Spring barley yield	ChC: ~82cM	3.56	0.39	0.01	0.00	0.10
Winter barley yield	ChC: ~85cM	4.70	0.4	0.11	0.02	0.14
Winter barley yield	ChE: ~56cM	4.75	0.4	0.11	0.01	0.13
Winter barley yield	ChG: ~26cM	4.68	0.4	0.11	0.03	0.16
Winter barley yield	ChG: ~55cM	3.06	0.4	0.11	0.01	0.09
Winter barley yield	ChF: ~117cM	3.10	0.4	0.11	0.00	0.04
Winter and spring barley yield	ChA: ~96cM	26.0	0.66	0.07	0.03	0.47
Winter and spring barley yield	ChC: 50-82cM	3.20	0.66	0.07	0.00	0.03
Winter and spring barley yield	ChG: ~117cM	16.75	0.66	0.07	0.02	0.23
Winter and spring barley yield	ChF: 150-170cM	4.47	0.66	0.07	0.01	0.13
Winter and spring barley yield	ChD: ~105cM	3.75	0.66	0.07	0.00	0.04

Detection of significant association is clearly a pre-requisite for the confirmation of QTL or the identification of genes and causative polymorphisms. It was never the primary intention that this study should go very far beyond investigating the practicality of association mapping in these species using historical data. However, the strength of association with the botanical traits led to additional work in which a gene and causative polymorphism have been identified (anthocyanin trait). Several other likely candidate genes (Table 8-6) are also noted for botanical traits in barley; one has since been corroborated (*int-c*: Ramsay *et al.* 2011), the others are all potential subjects for further research.

Other confirmation of associations can be obtained indirectly by testing the hypothesis that association discoveries are distributed randomly with respect to previously published QTL. Discoveries made for HFN, protein and yield in wheat appear to be distributed non-randomly relative to a sample of published QTL for these traits (7.6). In contrast the coincidence of barley yield and height associations with published QTL is not inconsistent with a random

distribution of discoveries. At first sight this appears paradoxical; both experiments give results consistent with power simulation yet the lower powered experiment produces discoveries which show a non-random spatial distribution; the more powerful experiment performed better but faild to show the same spatial enrichment. The explanation may lie in the genome sizes; the barley genome is considerably smaller (1200cM or 5.3gbp) than wheat (2500cM or 16gbp) and the 14 independent QTL for yield with average size 25cM cover about 30% of the genome while for plant height about 40% is covered. The consequence of high coverage is that association discoveries can easily fall into published QTL by chance, this means relatively large number of association discoveries would be necessary before the test had the power to detect deviation from the null. Since these are quantitative traits and power is known to be limited at low effect sizes the observed result is not surprising. The converse argument, that wheat was too successful given the panel size, can be similarly answered. The amount of genome covered by published QTL is smaller and the number of associated markers discovered is low making co-incidence less likely and hence giving more power to detect the deviation from the null.

Objectively, the failure to show a non-random distribution in the discovery of associations for barley height and yield is not evidence against the discoveries being genuine. Indeed, the success with other traits suggests some degree of optimism would be warranted if any of the identified markers were to be further evaluated.

9.5 Feasibility of association mapping in wheat and barley.

Moving to the central question of feasibility of association mapping in these species; the literature now contains many examples of studies in which association with trait has been reported in agricultural crops. However, the use of historically recorded phenotypes is less common. Also uncommon is any attempt to critically validate the discoveries against published QTL. The commonest approach has been a statement that some of the association discoveries were coincident with published QTL, while lacking an estimate of how likely that observation was if association discoveries are distributed randomly with respect to published QTL. Indeed, it is possible in the case of Roy *et al.* 2010 (4.1.1) to use the data presented by the authors to test this hypothesis and show that the success rate reported could reasonably be attributed to chance. Reimer *et al.* 2008 (4.1.4) note that a high proportion of discoveries are coincident with published QTL but in this case the proportion of markers in the panel which fall into published QTL is similarly high so no enrichment has been demonstrated.

The wheat and barley studies have, therefore, been more successful than many so far published probably because the panel sizes are larger and in the case of barley many of the traits are controlled by few loci of large effect.

9.6 Experimental design to increase power in wheat and barley association mapping.

It is evident that within the two panels association with yield and other quantitative traits is weak which has been shown to be an issue of statistical power. Extrapolation in simulated genotypes to larger panel sizes shows that power to detect QTL accounting about 10% of phenotypic variation could be as high as 80% if the variety and marker panels were increased to 1200 and 1500 respectively (6.4.2). To obtain samples from 1500 established varieties with corresponding phenotype is probably impossible given the constraints that varieties must have been officially registered before the genetic material enters the public domain, the desirability for varieties to be adapted to a common geographical region and also that sufficient trial information must be available to make a meaningful estimate of genetic effect. Marker panel size, however, is unlikely to be limiting as the number of wheat SNPs reported increases and genotyping chips are being developed to exploit them Akhunov *et.al.* (2009).

There are efforts being made to obtain large association panels in wheat (and other species[1]) using a multi-parent advanced generation inter-cross (MAGIC) (Cavanagh *et.al.* 2008). Developing from the proposals of Darvasi and Soller (1995), a parental generation typically comprising 4 or 8 varieties (inbred lines in the case of wheat) selected to represent maximum genetic diversity or perhaps to sample several important breeding pools. Inter-crossing between these can produce large numbers of highly re-combined individuals in an unstructured population, ideal for association mapping (see; Figure 9-2).

[1] Arabadopsis: Kover *et al.* 2009;
Rice and Sorghum http://www.generationcp.org/gcp_principal_investigators__

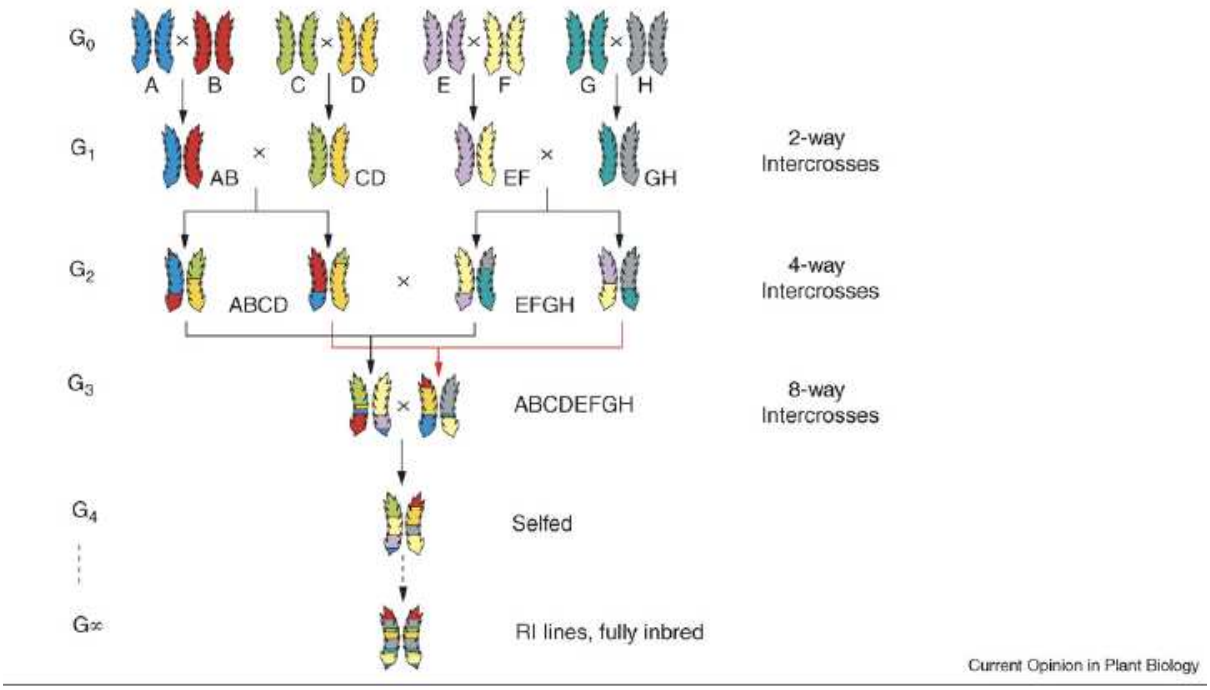


Figure 9-2. Multi-parent advance inter-cross (MAGIC). In this example eight genetically diverse fully inbred parental lines are inter-crossed over three generations to produce an eight-way inter-cross. The inter-cross population is then allowed to self-pollinate over multiple generations to produce a population of recombinant inbred lines. Multiple recombinations occur during this process and the final population can be almost any size. This means that an association study conducted on the final population will potentially benefit from both high power and high resolution. Reproduced from Cavanagh *et al.,* 2008).

Another approach to producing moderately unstructured and recombined populations is nested association mapping (NAM). As an example, Yu *et al.* 2008 and Myles *et al.* 2009 report a population in which 25 genetically diverse parental maize lines are genotyped at high resolution and then crossed with a reference inbred maize line (also genotyped). From the progeny recombinant inbred lines (RIL) are produced (5000 were planned). These moderately recombined RIL may be phenotyped and subjected to low resolution genotyping. Because the recombination is limited (essentially the equivalent of 2 meioses occur when a RIL is produced from an F1) the high resolution genotypes may be projected onto the low resolution genotypes of the RIL giving high resolution genotypes for 5000 lines at low cost. Although recombination is less extensive in this approach than in MAGIC the cost saving in genotyping should allow panels so large that power to detect associations will be high.

Nordborg and Weigel (2008) see NAM and MAGIC as conceptually very similar and note that such panels are essentially exploiting the strengths of bi-parental mapping and

association mapping in that there is in-built control of structure/kinship combined with large panels of (potentially) highly recombined individuals.

If structure is considered to relate to events distant in time then the NAM populations might be expected to exhibit some structural signature since each of the parental lines contributes to only one lineage; definition of this structural element in a regression model would be unambiguous; each lineage could be a co-factor. However, it would be reasonable to argue that the RIL within each NAM lineage were more closely related than between lineages; this might be modelled as a kinship matrix based on the pedigree or, given the extensive genotyping which is planned, using marker genotype. The kinship matrix from such analyses would reflect the lineages and its main feature would therefore be what might otherwise have been modelled as structure. It seems to me that NAM populations should be analysed for association using a mixed model approach including kinship as a random effect. This approach would have the advantage of recording the gradation of relatedness between RIL within a lineage. In simulation, however, Yu *et al.* 2008 find that correction for structural effects is unnecessary and they do not investigate the effect of kinship, however, the authors conclude that these simulations do not represent a general solution and the best analytical approach for each panel should be determined on a case by case basis.

MAGIC should avoid any signature from structure since the inter-crossing means each RIL samples all progenitor lines equally. Nonetheless Mendelian sampling will result in a gradation of kinship amongst the RIL which cannot be predicted from pedigree and which could confound an association study. In this case, again, I would suggest that mixed effects modelling with kinship as a random effect should be considered as an analytical strategy for association mapping.

9.7 Beyond association mapping.

Association mapping coupled with these artificial populations offers to plant geneticists considerable promise of well powered experiments not bedevilled by spurious associations. This, perhaps, will be the character of genetic mapping in plant science over the next five to ten years. However, increasing marker densities in genotyping platforms and ultimately re-sequencing of multiple individuals is already suggesting new analytical strategies. One such is genomic selection (Meuwissen *et al.* 2001).

Genomic selection is not strictly a mapping tool, however, it is a direct competitor to genetic mapping in that it has the potential to predict phenotype from genotype and so allow advance selection, which is the main practical application of all trait mapping in crop species.

Genomic selection predicts the phenotype of an individual for a quantitative trait using a regression model which uniformly samples the entire genome. Each segment of the genome has an effect size and a statistical significance and early publications tend to select the more significant segments and use them to develop a regression model to predict phenotype. The model is developed in a 'teaching' set and then applied to make advance predictions of the phenotypes of other individuals. Latterly there seems to be a trend towards using all available markers irrespective of significance.

The advantage of genomic selection, and advance selection in general, is speed in breeding programs. For example, genomic selection can be used to predict the potential milk yield or udder depth of a bull calf (VanRaden *et al.* 2009). This example seems perverse but this is simply breeding potential and has been used for many years to determine breeding strategies in animal and plant systems – crudely crossing the 'best with the second best and selecting the best'. The advantage of genomic selection is that it does not require breeding potential to be inferred retrospectively from measurements on mature adult animals or large plots of plants; genomic selection can be applied to seeds or embryos potentially reducing the cost of selection or (if more genotypes can be evaluated) increasing the intensity of selection. Both of which strategies have positive implications in a commercial situation.

Genomic selection relies on regression models with more independent variables than there are individuals in the teaching set. The obvious problem of over-estimation of effects may be dealt with by a penalised regression or Bayesian regression (see; Meuwissen *et al.* 2001; Hayes 2007).

Efforts are being made to apply genomic selection to plant populations but work at NIAB, in which I collaborated, suggests that current marker densities in wheat mean that the genomic selection model is predicting trait from kinship rather than the sum of small effects over the genome. Thus genomic selection could become a surrogate for pedigree breeding (a breeding strategy in which the performance of progeny is predicted in advance from the known performance of their progenitors) if marker densities are low. This result relates to the observation (Figure 2-6) that contemporaneous varieties are more closely related than varieties with greater age difference between them coupled to the upward trend in yield due

to genetic improvement (Silvey 1986, Mackay *et al.* 2011) over time. Essentially the model is predicting yield from a secondary correlation with age of variety.

In the development of genomic selection, MAGIC populations offer a useful test-bed since the RIL may be sampled randomly to obtain teaching and prediction sets which are not differentiated by kinship; alternatively RIL might be allocated to the two classes on the basis of common descent from individuals at G3 (**Figure 9-2**) thus introducing a potential differentiation between teaching and prediction sets based on kinship. Comparison between the two approaches would give some insight as to the changing mechanisms driving prediction as the MAGIC populations are genotyped more and more extensively over time. Similar potential is offered by the NAM populations in which the structural element presents another variable for study.

10 Conclusion

I conclude that despite the current limits on panel size imposed by numbers of varieties, available data and genotyping technology, it is possible to conduct successful association studies in wheat and barley using historical phenotypes and germplasm from seed banks. Structure/kinship imposed by breeding activity lead to very strong confounding signals which can swamp genuine associations in many cases. These effects may be controlled by various regression techniques but the most consistently effective is mixed effects modelling. The use of a kinship matrix based on correlation is a useful way to summarise both kinship and structure in the same model.

Power is limited in the wheat panel and discoveries of QTL for the quantitative traits tested may owe something to the 'winners curse'. However the discoveries are not randomly distributed in the genetic map when compared with QTL discovered by linkage analysis. Power simulations suggest that panel sizes of 1200 varieties and 1500 markers will give 80% power to detect loci individually contributing 10% of phenotypic variation.

Power was higher in the barley panel and the phenotypes included some which proved to be Mendelian; both in simulation and empirically the power to detect these was very high (>90%). QTL for quantitative traits may have been detected but the small genome means it is not possible to demonstrate that the distribution of discoveries was non-random relative to published QTL.

The locations found for QTL for the barley traits would all warrant further study, since as demonstrated with the anthocyanin trait, the barley panel is giving quite precise locations for the major association peaks.

The route I followed, using existing varieties and historic phenotypes, has proved useful and offered ready access to the ingredients necessary for association analysis. Unfortunately, the short period that most varieties spend in trial (≤ 2 years) makes much of the available phenotypic data insufficient for use, this coupled with the problem that un-registered varieties remain the private property of breeders means that alternative sources of germplasm will be needed if association mapping is to fulfil its potential. MAGIC and/or NAM populations offer a clear next step for wheat association mapping and at least one MAGIC wheat population is now approaching maturity. There will be a substantial phenotyping cost but beyond that the population will offer a developing resource since the RIL are essentially immortal and the phenotype, once recorded, is valid indefinitely so all future developments in genotyping may be applied retrospectively. Predicted developments in statistical genetics, such as the use of genomic selection, will also benefit from this resource.

The barley population was more successful because the smaller, simpler diploid genome had allowed higher marker density and the diverse barley lines shared a common set of botanical descriptors which were largely independent of environment. For the future, however, complex quantitative traits may need larger variety and marker panels; a MAGIC population in barley is also in development by SAC (Scottish Agricultural Colleges) although this is somewhat behind the situation in wheat.

I see association mapping in major crops being important for the next five to ten years but ultimately it will be superseded by genomic selection or other techniques taking advantage of increasing genotyping resolution or even re-sequencing.

In human genetics extensive local re-sequencing of individuals exhibiting extreme phenotypes is being used to uncover potentially causative SNPs within QTL/genes identified by GWAS (for example; Matsukawa *et al.* 2011; Khor and Li-Meng Goh, 2010). In fact the method described in this thesis to uncover the causative polymorphism for anthocyanin competence in barley is identical to re-sequencing of extreme phenotypes since in this case the phenotype naturally divided the panel in two extremes; the only material difference was in the amount of heritability. One can envisage that cheaper re-sequencing (perhaps confined to the transcriptome) and increasing computing capacity could well lead to the technique

being applied at a whole genome level as a method of primary discovery rather than confirmation. It is interesting to observe that such a development would, in essence, be little more than a bulked segregant analysis (Michelmore *et al.* 1991). Whilst crop genetics will inevitably lag behind human genetics in re-sequencing technology (because investment is spread over multiple species) the sequenced genome of rice is ripe for exploitation and synteny with other cereals may yield early benefits in unsequenced genomes.

However, there are many crop species which lack the resources afforded to wheat, maize and rice. Thus the techniques of MAGIC or NAM with suitable medium resolution genotyping could well allow association mapping to persist as a cost-effective tool for plant breeding in minor crops for many years to come.

11 References.

Akbari, M., Wenzl, P. Caig, V., Carling, J., Xia, L., Yang, S., Uszynski, G., Mohler, V., Lehmensiek, A., Kuchel, H., Hayden, M.J., Howes, N., Sharp, P., Vaughan, P., Rathmell, B., Huttner, E., Kilian, A. (2006). Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. Theor Appl Genet. 113:1409–1420

Akhunov, E., Nicolet, C. and Dvorak, J. (2009) Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. Theor Appl Genet. 119(3): 507–517.

Amasino, R.M. (2005) Vernalization and flowering time. Current Opinion in Biotechnology, 16:154–158.

Andersen, J. R., Zein, I., Wenzel, G., Krützfeldt, B., Eder, J., Ouzunova, M., and Lübberstedt, T. 2007. High levels of linkage disequilibrium and associations with forage quality at a *Phenylalanine Ammonia-Lyase* locus in European maize (*Zea mays* L.) inbreds. Theor Appl Genet 114, 307-319.

Aranzana MJ, Kim S, Zhao K, Bakker E, Horton M, *et al*. (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genet 1: 531-539.

Astle, W., and Balding, D.J., Population Structure and Cryptic Relatedness in Genetic Association Studies. 2009. Statistical Science. 24: 451-471.

Atwell, S. and multiple others. (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature, 465:627-631

Aulchenko, Y.S., Heutink, P., Mackay, I., Bertoli-Avella, A.M., Pullen, J., Vaessen1, N., Rademaker, T.A.M., Sandkuij, L.W. Cardon. L., Oostra, B., and van Duijn, C.M. (2004). Linkage disequilibrium in young genetically isolated Dutch population. European Journal of Human Genetics. 12: 527–534

Batey, I.L., Hayden, M.J., Cai, S., Sharp, P.J., Cornish, G.B., Morell, M.K. and Appels, R. (2001). Genetic mapping of commercially significant starch characteristics in wheat crosses. Aust. J. Agric. Res., 52, 1287–1296.

Beavis, W. D., 1998 QTL analyses: power, precision, and accuracy, pp. 145–162 in *Molecular Dissection of Complex Traits,* edited by A. H. Paterson. CRC Press, New York.

Beló, A., Zheng, P., Luck, S., Shen, B., Meyer, D.J., Li, B., Tingey, S. and Rafalski, A. (2008). Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. Mol Genet Genomics (2008) 279:1–10

Benjamini,Y. and Hochberg, Y (1995)Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Statistical Methodology).* 57: 289-300.

Blanco, A., Pasqualone, A., Troccoli, A., Di Fonzo, N., and Simeone, R. (2002). Detection of

grain protein content QTLs across environments in tetraploid wheats. Plant Molecular Biology 48: 615–623.

Bolibok-Brągoszewska, H., Heller-Uszyńska, K,. Wenzl, P., Uszyński, G., Kilian, A., and Rakoczy-Trojanowska, M. (2009) DArT markers for the rye genome - genetic diversity and mapping BMC Genomics , 10:578.

Bonjean AP, Angus WJ, 2001. The world wheat book. A history of wheat breeding. Paris: Laroisier Publishing.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. (2007) TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**:2633-2635.

Breseghello, F., and Sorrells, M.E., (2006). Association Mapping of Kernel Size and Milling Quality in Wheat (Triticum aestivum L.) Cultivars. Genetics 172: 1165–1177.

Campana, M.G., Hunt, H.V., Jones, H. and White, J. 2011. CorrSieve: software for summarizing and evaluating Structure Output. Molecular Ecology Resources. 11: 349-352

Campbell, K.G., Finney, P.L., Bergman, C.J., Gualberto, D.G., Anderson, J.A., Giroux, M.J., Siritunga, D., Zhu, J., Gendre, F., Roué, C., Vérel, A. and Sorrells, M.E. (2001).Quantitative Trait Loci Associated with Milling and Baking Quality in a Soft x Hard Wheat Cross. Crop Science. 41: 1275-1285

Cavanagh, C., Morell, M., Mackay, I. and Powell, W. 2008. From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Current Opinion in Plant Biology. 11:215–221

Chao, S., Zhang, W., Dubcovsky, J., and Sorrells, M,. (2007). Evaluation of Genetic Diversity and Genome-wide Linkage Disequilibrium among U.S. Wheat (*Triticum aestivum* L.) Germplasm Representing Different Market Classes Crop Sci. 47:1018–1030

Churchill, G.A. and Doerge, R.W. (1994) Empirical Threshold Values for Quantitative Trait Mapping. Genetics. 138: 963-971.

Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C., Ovington, N. R., Stevens, H. E., Nutland, S., Howson, J. M. M., Faham, M., Moorhead, M., Jones, H. B., Falkowski, M., Hardenbol, P., Willis, T. D., Todd, J. A., 2005. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37: 1243-6.

Cockram, J., White, J., Leigh, F.J, Lea, V.J., Chiapparino, E., Laurie, D.A., Mackay, I,J., Powell, W., and O'Sullivan, D.M. Association mapping of partitioning loci in barley. BMC Genetics 2008, 9:16

Cockram, J., White, J., Zuluaga[a], D.L., Smith, D., Comadran, J., Macaulay, M., Luo, Z., Kearsey, M.J., Werner, P., Harrap, D., Tapsell, C., Liu, H., Hedley, P.E., Stein, N., Schulte, D., Steuernagel, B., Marshall, D.F., Thomas, W.T.B., Ramsay, L., Mackay, I., Balding, D.J., The AGOUEB Consortium, Waugh, R. and Donal M. O'Sullivan, D.M. 2010. Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. PNAS. 107: 21611-21616

Cover, T., and P. Hart, 1967 Nearest neighbor pattern classification. Proc IEEE Trans Inform Theory 13.

Crossa, J., Burgueño, J., Dreisigacker, S., Vargas, M., Herrera-Foessel, S..A., Lillemo, M., Singh, R.P., Trethowan, R., Warburton, M., Franco, J., Reynolds, M., Crouch, J.H., and Ortiz, R. (2007). Association Analysis of Historical Bread Wheat Germplasm Using Additive Genetic Covariance of Relatives and Population Structure. Genetics 177: 1889–1913.

D'hoop, B.B., Paulo, M.J., Mank, R.A., van Eck, K.J. and van Eeuwijk, F.A. (2008) Association mapping of quality traits in potato (*Solanum tuberosum* L.). Euphytica 161:47–60

Darvasi, A and Soller, M. (1995) Advanced intercross lines, an experimental population for fine genetic mapping. Genetics. 141: 1199-1207.

Dawson, E., Abecasis, G.R., *et al*. (2002). A first-generation linkage disequilibrium map of human chromosome 22. Nature. Vol 418. 544-548.

Devlin, B. and Roeder, K. (1999) Genomic Control for Association Studies. Biometrics. 55: 997-1004

Devos, K.M. and Gale, M.D. (1997) Comparative genetics in the grasses. Plant Molecular Biology 35: 3–15.

Dhaliwal, A.S., D.J. Mares, and D.R. Marshall. 1987. Effect of 1B/ 1R chromosome translocation on milling and quality characteristics of bread wheats. Cereal Chem. 64:72–76

Distelfield, A., Korol, A., Dubcovsky, J., Uauy, C., Blake, T. and Fahima, T. (2008). Colinearity between the barley grain protein content (GPC) QTL on chromosome arm 6HS and the wheat Gpc-B1 region. Molecular Breeding. 22: 25-38

Distelfield, A., Uauy, C., Fahima, T. and Dubcovsky, J. (2006). Physical map of the wheat high-grain protein content gene Gpc-B1 and development of a high throughput molecular marker. New Phytologist. 169: 753-763.

Ellis, M.H., Spielmeyer, W., · Gale, K.R., Rebetzke, G.J. and · Richards, R.A. (2002) "Perfect" markers for the Rht-B1b and Rht-1b dwarfing genes in wheat Theor Appl Genet. 105: 1038–1042.

Engledow, F.L. (1920) Inheritance in barley. I The lateral florets and the rachilla. Journal of Genetics. 10:2. 93-108.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software Structure: a simulation study. Molecular Ecology. 14:2611–2620.

Excoffier, L. G. Laval, and S. Schneider (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1:47-50.

Excoffier, L., Smouse, P.E. and Quattro, J.M. (1992) Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. Genetics 131: 479-491

Falush, D., Stephens, M. and Pritchard, J.K. (2003) Inference of population structure from

multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164:1567-1587.

Falush, D., Stephens, M. and Pritchard, J.K. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Molecular Ecology Notes (2007) doi: 10.1111/j.1471-8286.2007.01758.x (Available online).

FAO (2005). Food and Agriculture Organisation of the United Nations. Production Statistics. http://faostat.fao.org/

Fellay, J., Shianna, K.V., Ge, D., Colombo, S., Ledergerber, B., Weale, B., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., Cozzi-Lepri, A., De Luca, A., Easterbrook, P., Francioli, P., Mallal, S., Martinez-Picado, ., Miro, J.M., Obel, N., Smith, J.P., Wyniger, J., Descombes, P., Antonarakis, S.E., Letvin, N.L., McMichael, A.J.,Haynes, B.F.,Telenti, A., Goldstein1, D.B. 2007. A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. Vol 317: 944:947

Flintham, J.E., Angus, W.J., Gale, M.D., (1997) Journal of Agricultural Science, Cambridge (1997), 129: 371-378.

Galwey, N.W. (2006) Introduction to Mixed Modelling. Publ: John Wiley & Sons Ltd., Chichester.

Galwey, N.W. (2011). Pers Comm.

Giles, R.J. (1989). The frequency of natural cross-fertilisation in sequential sowings of winter Barley. Euphytica 43:125-134.

Gill, B.S., Appels, R., Botha-Oberholster, A-M., Buell, C.R., Bennetzen, J.L., Chalhoub, B., Chumley, F., Dvorˇa´k, J., Iwanaga, M., Keller, B., Li, W., McCombie, W.R., Ogihara, Y., Quetier. F. and Sasaki. T. (2004) A Workshop Report on Wheat Genome Sequencing: International Genome Research on Wheat Consortium. Genetics 168: 1087–1096

Gordon D., Finch S.J., Nothnagel M., and Ott J. (2002) Power and Sample Size Calculations for Case-Control Genetic Association Tests when Errors Are Present: Application to Single

Nucleotide Polymorphisms. Human Heredity 54:22-33

Gordon, D., Finch, S.J., Nothnagel, M. and Otta, J. 2002. Power and Sample Size Calculations for Case-Control Genetic Association Tests when Errors Are Present: Application to Single Nucleotide Polymorphisms. Hum Hered. 54:22–33 .

Groos, C., Gay, G., Perretant, M-R., Gervais, L., Bernard, M., Dedryver, F. and Charmet, G. (2002). Study of the relationship between pre-harvest sprouting and grain color by quantitative trait loci analysis in a white×red grain bread-wheat cross. Theor Appl Genet. 104: 39–47.

Groos, C., Robert, N., Bervas, E. and Charmet, G. (2003). Genetic analysis of grain protein-content,grain yield and thousand-kernel weight in bread wheat. Theor Appl Genet. 106:1032–1040

Hamblin, M.T., Close, T.J., Bhat, P.R., Chao, S., Kling, J.G., Abraham, K.J., Blake, T., Brooks, W.S., Cooper, B., Griffey, C.A., Hayes, P.M., Hole, D.J., Horsley, R.D., Obert, D.E.,

Smith, K.P., Ullrich, S.E., Muehlbauer, G.J. and Jannink, J-L. (2010) Population Structure and Linkage Disequilibrium in U.S. Barley Germplasm: Implications for Association Mapping. Crop Science. 50:556-566.

Haseneyer, G., Stracke,S,, Paul, C., Einfeldt, C., Broda, A., Piepho, H.-P., Graner, A., and Geiger, H.H. (2010) Population structure and phenotypic variation of a spring barley world collection set up for association studies. Plant Breeding. 129:271-279

Hayes, B. (2007) QTL Mapping, MAS, and Genomic Selection. Course notes, download from:
http://www.anslab.iastate.edu/Class/ABG%20Short%20Course/QTL%20Mapping,%20MAS,%20and%20Genomic%20Selection%20Dr.%20Ben%20Hayes.pdf

Hom, G., Graham, R.R., Modrek, B., Taylor, K.E., Ortmann, W., Garnier, S., Lee, A.T., Chung, S.A. Ferreira, R.C., Pant, K., Ballinger, D.G., Kosoy, R., Demirci, Y., Kamboh, M.I., Kao, A.H., Tian, C., Gunnarsson, I., Bengtsson, A.A., Rantapää-Dahlqvist, S., Petri, M., Manzi, S., Seldin, M.F., Rönnblom, L., Syvänen, A., Criswell, L.A., Gregersen, P.K., Behrens, T.W. 2008. Association of Systemic Lupus Erythematosus

with *C8orf13–BLK* and *ITGAM–ITGAX*. N Engl J Med 2008;358:900-9.

Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A., Riley, J., Purvis, I., and Xu, C-F., 2004. Detection of genotyping errors by Hardy–Weinberg equilibrium testing. European Journal of Human Genetics. 12: 395–399.

Huang, X.Q., Kempf, H., Ganal, M.W. and Röder, M.S. 2004. Advanced backcross QTL analysis in progenies derived from a cross between a German elite winter wheat variety and a synthetic wheat (Triticum aestivumL.). Theor Appl Genet. 109: 933–943

Hudson, R. R. (2002) Generating samples under a Wright-Fisher neutral model. Bioinformatics 18:337-8,

Jaccoud, D., Peng, K., Feinstein, D. and Kilian, A. (2001) Diversity Arrays: a solid state technology for sequence information independent genotyping. Nucleic Acid Research. Vol 29. No 4. e25.

Jordan, W.R., Duga, W.A. and Shouse, P.J. 1983. Strategies for crop improvement for drought-prone regions. Agricultural Water Management. 7: 281-299.

Kang, M.K., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J. and Eskin, E. (2008) Efficient control of population structure in model organism association mapping. Genetics. 178:1709-1723.

Kato, K., Miura, H. and Sawada, S. (2000). Mapping QTLs controlling grain yield and its components on chromosome 5A of wheat. Theoretical and applied genetics. 101: 1114-1121

Khor, C.C. and Li-Meng Goh, D. 2010. Strategies for identifying the genetic basis of dyslipidemia: genome-wide association studies vs. the resequencing of extremes. Current Opinion in Lipidology, 21:123–127

Kilian, A. (2001). Diversity Array Technology. World Patent No. WO 01/73119 A2

Kosambi, D.D., 1944. The estimation of map distances from recombination values. Ann. Eug 12 : 172-175.

Kover, P.X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I.M., Purugganan, M.D., Durrant, C. And Mott, R. (2009) A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in Arabidopsis thaliana. PLoS. Volume 5 | Issue 7 (Online).

Kraakman, At.T.W., Niks, R.E., Van den Berg, P.M.M.M., Van Eeuwijk, P.S. and F.A. (2004) Linkage Disequilibrium Mapping of Yield and Yield Stability in Modern Spring Barley Cultivars. Genetics 168: 435–446.

Kulwal, P.L., Singh, R., Balyan, H.S. and Gupta, P.K. (2004), Genetic basis of pre-harvest sprouting tolerance using single-locus and two-locus QTL analyses in bread wheat. Funct Integr Genomics. 4: 94–101

Kunert, A., Naz, A.A., Dedeck, O., Pillen, K. and Léon, J. (2007). AB-QTL analysis in winter wheat: I. Synthetic hexaploid wheat (*T. turgidum ssp. Dicoccoides x T. tauschii*) as a source of favourable alleles for milling and baking quality traits. Theor Appl Genet. 115: 683–695.

Li, C., Ni, P., Francki, M., Hunter, A., Zhang, Y., Schibecci, D., Li, H., Tarr, A., Wang, J., Cakir, M., Yu, J., Bellgard, M., Lance, R. and Appels, R. (2004).Genes controlling seed dormancy and pre-harvest sprouting in a rise-wheat-barley comparison. Funct. Integr. Genomics. 4: 84-93.

Li, L., Paulo, M-J., Strahwald, J., Lübeck, J., Hofferbert, H.R., Tacke, E., Junghans, H., Wunder, J., Draffehn, S., van Eeuwijk, F.A., and Gebhardt, C. (2008). Natural DNA variation at candidate loci is associated with potato chip color, tuber starch content, yield and starch yield. Theor Appl Genet 116:1167–1181.

Li, L., Paulo, M-J., van Eeuwijk, F., and Gebhardt, C. (2010). Statistical epistasis between candidate gene alleles for complex tuber traits in an association mapping population of tetraploid potato. Theor Appl Genet 121:1303–1310

Li, S., Jia, J., Wei, X., Zhang, X., Li, L., Chen, H., Fan, Y., Sun, H., Zhao, X., Lei, T., Xu, Y., Jiang, F., Wang, H. and Li, L. (2007). A intervarietal genetic map and QTL analysis for yield traits in wheat. Mol Breeding (2007) 20:167–178

Linkiewicz, A.M., L. L. Qi,,1 B. S. Gill, A. Ratnasiri, B. Echalier, S. Chao, G. R. Lazo, D. D. Hummel, O. D. Anderson, E. D. Akhunov, J. Dvorˇa´k, M. S. Pathan, H. T. Nguyen, J. H. Peng, N. L. V. Lapitan, Miftahudin, J. P. Gustafson, C. M. La Rota, M. E. Sorrells, K. G. Hossain, V. Kalavacharla, S. F. Kianian, D. Sandhu, S. N. Bondareva, K. S. Gill, E. J. Conley, J. A. Anderson (2004). A 2500-Locus Bin Map of Wheat Homoeologous Group 5 Provides Insights on Gene Distribution and Colinearity With Rice. Genetics 168: 665–676

Loiselle, B.A., Sork, V.L., Nason, J. and Graham, C. 1995. Spatial genetic structure of a tropical understory shrub, *Psychotria officianalis* (Rumiacea). American Journal of Botany. 82: 1420-1425.

Lynch, M., (1988) Estimation of relatedness by DNA Fingerprinting. Mol. Biol. Evol. 5: 584-599.

Maccaferri, M., Sanguineti, M. C., Noli, E., and Tuberosa, R. 2005. Population structure and long-range linkage disequilibrium in a durum wheat elite collection. Molecular Breeding 15, 271-289.

Maccaferri, M., Sanguineti, M. C., Xie, C., Smith, J. S. C., and Tuberosa, R. 2007. Relationships among durum wheat accessions. II. A comparison of molecular and pedigree information. Genome 50, 385-399.

Maccaferri, M., Sanguineti, M.C., Demontis, A., El-Ahmed, A., del Moral, L.G., Maalouf, F., Nachit, M., Nserallah7, N., Ouabbou, H., Rhouma, S., Royo, C., Villegas, D. and Tuberosa, R. (2010) Association mapping in durum wheat grown across a broad range of water regimes Journal of Experimental Botany. doi:10.1093/jxb/erq287 (Online only at time of writing).

Maccaferri, M., Sanguineti, M.C., Mantovani, P., Demontis, A., Massi, A., Ammar, K., Kolmer, J.A., Czembor, J.H., Ezrati, S., Tuberosa, R. (2010a) Association mapping of leaf rust response in durum wheat. Mol Breeding. 26:189–228.

MacCluer, J.W., Vandeberg J.L., Read, B., Ryder, O.A (1986). Pedigree analysis by computer-simulation. Zoo Biology 5: 147-160.

Mackay, I., Horwell, A., Garner, J., White, J., McKee, J. and Philpott, H. 2011 Reanalyses of the historical series of UK variety trials to quantify the contributions of genetic and environmental factors to trends and variability in yield over time. Theoretical and Applied Genetics. 122: 225–238

Malecot, G. (1948) Les Mathematiques de l'Hérédité (Masson et Cie, Paris).

Malosetti, M., van der Linden, C.G., Vosman, B. and van Eeuwijk, F.A. (2007) A mixed model approach to association mapping using pedigree information with an illustration to resistance for phytophthora infestans in potato. Genetics, 175: 879-889.

Matsukawa, T., Asheuer, M., Takahashi, Y., Goto, J., Suzuki, Y., Shimozawa, N. and Takano, H. 2011. Identification of novel SNPs of ABCD1, ABCD2, ABCD3, and ABCD4 genes in patients with X-linked adrenoleukodystrophy (ALD) based on comprehensive resequencing and association studies with ALD phenotypes. Neurogenetics 12:41–50.

Matthies, I.E., Weise, S., Förster, J., Röder, M.S. 2009 Association mapping and marker development of the candidate genes (1 →3),(1 → 4)-β-D-Glucan-4-glucanohydrolase and (1 → 4)-β-Xylan-endohydrolase 1 for malting quality in barley. Euphytica 170:109–122

McCartney, C.A., Somers, D.J., Humphreys, D.G., Lukow, O., Ames, N., Noll, J., Cloutier, S., and McCallum, B.D. (2005). Mapping quantitative trait loci controlling agronomic traits in the spring wheat cross RL4452 x 'AC Domain'. Genome. 48: 870-883

Melchinger, A. E., M. M. Messmer, M. Lee, W. L. Woodman, and K. R. Lamkey, (1991) Diversity and relationships among U.S. maize inbreds revealed by restriction fragment length polymorphisms. Crop Sci. 31: 669-678.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics. 157: 1819–1829

Michelmore, R.W., Paran, I. and Kesseli, R.V. 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. Proc. Natl. Acad. Sci. Vol. 88, pp. 9828-9832.

Miedaner, T., Wu¨rschum, T., Maurer, H.P., Korzun, V., Ebmeyer, E. and Reif, J.C. Association mapping for Fusarium head blight resistance in European soft winter wheat. 2010. Mol Breeding. DOI 10.1007/s11032-010-9516-z. Online only at time of writing.

Myles, S., Peiffer, J., Brown, P.J., Ersoz, E.S., Zhang, Z., Costich, D.E. and Buckler, E.S. 2009. Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. The Plant Cell. 21: 2194–2202.

Neumann, K., Kobiljski, b., Denčić,S., Varshney, R.K. and Börner, A. (2010) Genome-wide association mapping: a case study in bread wheat (*Triticum aestivum L.*) Mol Breeding DOI 10.1007/s11032-010-9411-7 (Online only at time of writing).

Nordborg, M., and Weigel, D. (2008) Next-generation genetics in plants. Nature. 456: 720-723.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2(12): e190.

Payne, R.W., Murray, D.A., Harding, S.A., Baird, D.B. & Soutar, D.M. (2009). *GenStat for Windows (12th Edition) Introduction*. VSN International, Hemel Hempstead.

Peng, J.H., Bai Y., Haley, S.D. and Lapitan, N.L.V. 2009 Microsatellite-based molecular diversity of bread wheat germplasm and association mapping of wheat resistance to the Russian wheat aphid. Genetica 135:95–122

Piepho, H.P. Möhring, J., Melchinger, A.E. and Büchse, A. (2008) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica. 161:209–228

Plenge., R.M., et al. (2007) Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nature Genetics. 39(12):1477-82.

Prasad, M., Kumar, N., Kulwal, P.L., Roder, M.S., Balyan, H.S., Dhaliwal, H.S. and Gupta, P.K. (2003). QTL analysis for grain protein content using SSR markers and validation studies using NILs in bread wheat. Theoretical and Applied Genetics. 106: 659-667.

Price, A..L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, R. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 38: 904-909.

Pritchard JK, Donnelly P. (2001) Case-control studies of association in structured or admixed populations. Theor Popul Biol. 60: 227-237.

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

Pritchard, J.K., Stephens,M., Rosenberg, N.A., and Donnelly, P. (2000a) Association

Mapping in Structured Populations Am. J. Hum. Genet. 67:170–181.

Purcell S, Cherny SS, Sham PC. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics, 19 (1):149-150.

Quarrie, S.A., Quarrie, S.P., Radosevic, R., Rancic, D., Kaminska, A., Barnes, J.D., Leverington, M., Ceoloni, C., and Dodig, D. (2006). Dissecting a wheat QTL for yield present in a range of environments: from the QTL to candidate genes. Journal of Experimental Botany, Vol. 57, No. 11, pp. 2627–2637

Quarrie, S.A., Steed, A., Calestani, C., Semikhodskii, A., Lebreton, C., Chinoy, C., Steele, N., Pljevljakusic´, D., Waterman, E., Weyen, J., Schondelmaier, J., Habash, D.Z., Farmer, P., Saker, L., Clarkson, D.T., Abugalieva, A., Yessimbekova, M., Turuspekov, Y., Abugalieva, S., Tuberosa, R., Sanguineti, M-C., Hollington, P.A., Aragues, R., Royo, A. and Dodig, D. (2005). A high-density genetic map of hexaploid wheat (Triticum aestivum L.) from the cross Chinese Spring x SQ1 and its use to compare QTLs for grain yield across a range of environments. Theor Appl Genet 110: 865–880

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ramsay, L., Comadran, J., Druka, D., Marshall, D.F., Thomas, W.T.B., Macaulay, M., MacKenzie, K., Simpson, C., Fuller, J., Bonar, N., Hayes, P.M., Lundqvist, U., Franckowiak, J.D., Close, T.J., Muehlbauer, **G.J. and** Waugh, **R. 2011** *INTERMEDIUM-C*, a modifier of lateral spikelet fertility in barley, is an ortholog of the maize domestication gene *TEOSINTE BRANCHED 1*. Nature Genetics. 43:2. 169-172.

Ravel, C., Praud, S., Murigneux, A., Linossier, L., Dardevet, M., Balfourier, F., Dufour, P., Brunel, D., and Charmet, G. 2006. Identification of *Glu-B1-1* as a candidate gene for the quantity of high-molecular-weight glutenin in bread wheat (Triticum aestivum L.) by means of an association study. Theor Appl Genet 112, 738-743.

Reimer, S., Pozniak, C.J., Clarke, F.R., Clarke, J.M., somers, D.J., Knox, R.E., and Singh, A.K. (2008) Association mapping of yellow pigment ina an elite collection of durum wheat cultivars and breeding lines. Gemome. 51:1016-1025

Remington, D.L., Thornsberry, J.M., Matsuoka, Y., Wilson, L.M., Whitt, S.R., Doebley, J., Kresovich, S., Goodman, M.M. and Buckler, E.S., (2001). Structure of Linkage Disequilibrium and Phenotypic Associations in the Maize Genome. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 98, No.20: 11479-11484.

Rostocks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, Graner A, Close TJ, Waugh R: Recent history of artificial outcrossing facilitates whole genome association mapping in elite crop varieties. Proc Natl Acad Sci USA 2006 , 103:18656-18661

Rostoks, N., Ramsay, L., MacKenzie, K., Cardle, L., Bhat, P.R., Roose, M.L., Svensson, J.T., Stein, N., Varshney, R.K., Marshall, F.F., Graner, A., Close, T.J., and Waugh, R. (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite

inbred crop varieties. PNAS. vol. 103. no. 49. 18656–18661

Roy, J.K., Smith, K.P., Muehlbauer, G.J., Chao, S., Close, T.J., Steffenson, B.J. (2010) Association mapping of spot blotch resistance in wild barley. Mol Breeding 26:243–256

Roy, J.K., Prasad, M., Varshney, R.K., Balyan, H.S., Blake, T.K., Dhaliwal, H.S., H-Singh, Edwards, K.J. and Gupta, P.K. (1999) Identification of a microsatellite on chromosomes 6B and a STS on 7D of bread wheat showing an association with preharvest sprouting tolerance. Theor Appl Genet. 99: 336-340

Rush, T. 2010. Cereals 2010: Plant breeders call for new approach to R&D funding. Farmer's Guardian. 9th June 2010.

Semagn, K., Bjornstad, Å., Skinnes, H., Marøy, A.G., Tarkegne, Y., Tarkegne, Y., and William, M. (2006). Distribution of DArT, AFLP and SSR markers in a genetic linkage map of a double haploid hexaploid wheat population. Genome 49: 545–555.

Setakis, E., Stirnadel, H. and Balding, D.J. (2006) Logistic regression protects against population structure in genetic association studies. Genome Res. 16: 290-296.

Silvey, V., (1986) The Contribution of New Varieties to Cereal Yields in England and Wales between 1947 and 1983, Journal of the National Institute of Agricultural Botany. 1: 155-168.

Snape, J.W., Foulkes, M.J., Simmonds, J., Leverington, M., Fish, L.J., Wang, Y., and Ciavarrella, M. (2007). Dissecting gene x environmental effects on wheat yields via QTL and physiological analysis. Euphytica (2007) 154:401–408.

Snedecor, G.W. & W.G. Cochran, 1973. Statistical Methods, 6th

Edn., Iowa State University Press, Iowa, USA

Somers, D.J., Banks, T., DePauw, R., Fox, S., Clarke, J., Pozniak, C. and McCartney, C. (2007) Genome-wide linkage disequilibrium analysis in bread wheat and durum wheat . Genome. 50: 557-567

Somers, D.J., Isaac, P., Edwards, K. (2004). A high-density microsatellite consensus map for bread wheat (Triticum aestivum L.). Theor Appl Genet 109: 1105–1114

Speilman, R.J. and Ewens, W.J., (1996). The TDT and Other Family-Based Tests for Linkage Disequilibrium and Association. American Journal of Human Genetics. 59:983-989.

Stich, B., Möhring, J., Piepho, H., Heckenberger, M.,.Buckler, E.S., and Melchinger, A.E. (2008). Comparison of Mixed-Model Approaches for Association Mapping. Genetics, Vol. 178: 1745-1754.

Stich, B., Piepho, H-P., Schulz, B., and Melchinger, A.E. (2008). Multi-trait association mapping in sugar beet (*Beta vulgaris* L.) Theor Appl Genet. 117(6):947-954

Storey, J.D. (2002) A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. 64: 479-498.

Storey, J.D. (2003) The Positive False Discovery Rate: a Bayesian Interpretation and the Q –

Value. *The Annals* of Statistics. 31: 2013-2035.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies PNAS. 100: 9440–9445.

Sun, D., Ren, W., Sun, G., Peng, J. (2010) Molecular diversity and association mapping of quantitative traits in Tibetan wild and worldwide originated barley (Hordeum vulgare L.) germplasm. Euphytica DOI: 10.1007/s10681-010-0260-6 (Online only at time of writing)

Sun, H., Lü, J., Fan, Y., Zhao, Y., Kong, F., Li, R., W, H. and Li, S. (2008). Quantitative trait loci (QTLs) for quality traits related to protein and starch in wheat. Progress in Natural Sciences 18: 825-831.

Sutka, J. (2001) Genes for frost resistance in wheat. Euphytica 119: 167–172.

Sutka, J. and Snape, J.W. (1989) Location of a gene for frost resistance on chromosome 5A of wheat. Euphytica 42: 1-44

Sweeney MT, Thomson MJ, Pfeil BE, McCouch S (2006) Caught red-handed: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. Plant Cell 18:283–294.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 447: 661-678.

Tommasini. L., Schnurbusch, T., Fossati, D., Mascher F., and Keller, B. (2007) Association mapping of Stagonospora nodorum blotch resistance in modern European winter wheat varieties. Theoretical and Applied Genetics. 115: 40-5752

UPOV (1991) INTERNATIONAL CONVENTION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS of December 2, 1961, as Revised at Geneva on November 10, 1972, on October 23, 1978, and on March 19, 1991. (http://www.upov.int/en/publications/conventions/1991/act1991.htm)

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., and Schenkel, F. (2009). *Invited review:* Reliability of genomic predictions

for North American Holstein bulls. *J. Dairy Sci.* 92:16-24.

Veyrieras, J-B., Goffinet, B., Charcosset, A. (2007). MetaQTL: a package of new computational methods for the meta-analysis of QTL mapping experiments. *BMC Bioinformatics* 2007, 8:49.

Waines JG, Ehdaie B. 2007. Domestication and Crop Physiology: Roots of green-revolution wheat. Annals of Botany 100:991-998.

Weber, A,.Clark, R.M., Vaughn, L., de Jesu´s Sa´nchez-Gonzalez, J., Yu, J., Yandell, B.S., Bradbury, P., and Doebley, J. (2007). Major Regulatory Genes in Maize Contribute to Standing Variation in Teosinte (*Zea mays* ssp. *parviglumis*). Genetics 177: 2349–2359

Welcome trust case control study (2007).Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. Vol 447:661-684.

Wellcome Trust Case Control Consortium (2007) Association scan of 14,500

nonsynonymous SNPs in four diseases identifies autoimmunity variants.
Nat Genet 39(11): 1329-1337

Wellington, P.S. and Silvey, V.S. 1997. New Variety Tests and Farmers' Leaflets. Chapter 5. pp 81-82. In: Crop and Seed Improvement. Publ, The National Institute of Agricultural Botany, Cambridge.

Wenzl, P., Li, H., Carling, J., Zhou, M., Raman, H., Paul, E., Hearnden, P., Maier, C., Xia, L., Caig. V., Ovesná, J., Cakir, M., Poulsen, D., Wang, J., Raman, R., Smith, K.P., Muehlbauer, G.J., Chalmers, K.J., Kleinhofs, D., Huttner, E., and Kilian, A. (2006) A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and agricultural traits. BMC Genomics, 7:206.

Westfall, P.H. and Young, S.S. (1993) On Adjusting P-Values for Multiplicity. Biometrics. 49: 941-945.

Wheat CRC (2007) http://www.wheat-research.com.au/hot-topics.html.

White, J. (2001). Recommending quality wheat varieties; Chasing a moving target. Aspects of Applied Biology. 64: 75 – 78

White, J., Law, J.R., MacKay, I., Chalmers, K.J., Smith, J.S.C., Kilian A. and. Powell W. (2008). The genetic diversity of UK, US and Australian cultivars of *Triticum aestivum* measured by DArT markers and considered by genome. Theoretical and Applied Genetics. Volume 116. Number 3. 439-453.

Wicker, T., Narechania, A., Sabot, F., Stein, J., Vu, G.T.H., Graner, A., Ware, D., and Stein N. (2008). Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *BMC Genomics* 2008, 9:518

Wright, S. (1943) Isolation by distance. Genetics. 38 : 114-138

Wright, S. (1951) The genetical structure of populations. Annals of Eugenics, 15, pp. 323-354.

Wright, S., 1951. The general structure of populations. An Eugen 15:323-354.

Wright, S., 1965. the interpretation of population structure by F-statistics with special regard to systems of mating. Evol 19:395-420.

Xu, J., Turner, A., Little, J., Bleecker e.r. and Meyers, D.A. 2002. Positive results in association studies are associated with departure from Hardy-Weinberg equilibrium: hint for genotyping error? Hum Genet. 111 :573–574.

Xu, S., (2003) Theoretical Basis of the Beavis Effect. Genetics 165: 2259–2268

Yao, J., Wang, L., Liu, L., Zhao, C. and Zheng, Z. (2009) Association mapping of agronomic traits on chromosome 2A of wheat. Genetica 137:67–75.

Yu, J., Holland, J.B., McMullen, M.D., and Buckler, E.S. 2008 Genetic Design and Statistical Power of Nested Association Mapping in Maize. Genetics. 178: 539–551.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D.,

Gaut, B.S., Holland, J.B., Kresovich, S., and Buckler, E.S. (2006). A Unified Mixed-Model Method for Association Mapping Accounting for Multiple Levels of Relatedness. Nature Genetics. 38: 203-208.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Holland, J.B., Kresovich, S., and Buckler, E.S. (2006). A Unified Mixed-Model Method for Association Mapping Accounting for Multiple Levels of Relatedness. Nature Genetics. 38: 203-208

Zanetti, S. Winzler, M., Keller, M., Keller, B.,Messmer, M. (2000) Genetic analysis of pre-harvest sprouting resistance in a wheat x spelt cross. Crop. Sci. 40:1406-1417.

Zhang, L.Y., Marchard, S., Tinker, N.A., Belzile, F. (2009) Population structure and linkage disequilibrium in barley assessed by DArT markers. Theor Appl Genet. 119:43–52

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, *et al*. (2007) An Arabidopsis example of association mapping in structured samples. PLoS Genet 3(1):71-82

.

.

12 Appendices.

12.1 Appendix 1. The mixed effects model (MM): a summary.

Mixed modelling will be used in this research in two ways; to estimate variety effects for traits from incomplete historical trials data and to define the structure of the genetic variance in an association panel which can be accounted in terms of pair-wise kinship over varieties. It is thus important to describe the general principles of mixed modelling to give a context in which to describe the models used in analysis.

The simplest example of a MM is a linear model containing at least two fixed and two random terms (at least two because the residual error is the first random term and the intercept is the first fixed term). The distinction between random and fixed effects is to a large extent defined by the objective of the experiment and some judgement is required; this is one reason why mixed effects modelling cannot be viewed as a completely routine operation. Broadly speaking, fixed terms are those factors defined by the researcher (for example; fertilizer rate, variety, drug dose, etc) and random terms are factors recognised as contributing to total variance but which have been sampled from all possible representations of those factors (for example; field blocks, geographical sites, harvesting machines).

Modelling with fixed terms.

In simple regression analysis the response variable has a value defined in terms of the measured values of the explanatory variables and their coefficients. There is a random term in the model – the residual error, *e*, which is assumed to be distributed as $e \sim N(0, \sigma_e^2)$, independently of all other variables.

The familiar simple linear model can be written:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

To give an example, suppose we attempt to estimate variety yield for varieties Alpha, Beta, Gamma and Delta in a yield trial in which each variety is present twice in each of four blocks (A,B,C and D). The simplest (fixed effects) model would be:

$$Y = \mu + Z_v \beta_v + e \qquad (1)$$

Where Y is a vector of observed yields, μ is the average yield over all plots $\beta_v$is a vector of variety effects, Z_v is an incidence matrix of varieties (number of columns = number of varieties=length of β_v; number of rows = number of observations; 1=variety present, 0 = variety absent from a given observation) and *e* is a vector of residual errors. Adjusting β to minimise the sum of squares of *e* gives the least squares solution to the regression model.

A more sophisticated fixed-effects model including block information would be:

$$Y = \mu + Z_v \beta_v + Z_b \beta_b + e \qquad (2)$$

The additional terms being Z_b an incidence matrix of blocks and β_b a vector of block effects. Indeed block and variety effects may be combined in a single vector of fixed effects with a corresponding incidence matrix.

The limitations of the fixed model.

The model (1) gives equal weight to all observations; it assumes they are independent given the covariates. However, there is the possibility of a nuisance variable tending to (for example) make plots in a given area of a field over or under-perform together. The purpose of blocking (model (2)) is to partition out the variation caused by these local effects, and thus decrease the dependence of the observations on residual variety effect.

Both these models give equal weight to all observations. This is not a problem if each variety is equally represented in all blocks but in reality an experiment might be unbalanced due to design, crop failure or human error and as a consequence varieties may be present different numbers of times in each block. In fixed effects models the missing data is dealt with by reducing the degrees of freedom appropriately and finding the least squares solution for the model as normal. The variety effects estimated from the model are correct for the data but extrapolation – for example estimating breeding value in future generations - is problematic if the effects for the varieties have all been estimated from different numbers of observations.

The MM.

Unbalanced experiments.

Equation (1) described a model in which the possibility that observations for the different varieties might not be independent was ignored. Equation (2) added block structure so that one source of co-variation was removed. There remains a problem with incomplete data, which leads to the risk of over-stating the magnitude of variety effects. The MM offers an alternate solution in which the estimated effect sizes are shrunk to reflect the uncertainty with which they are known. The model can be represented as:

$$Y = \mu + Z_v\beta_v + Z_bU_b + e \qquad (3)$$

where U_b is a vector of random effects of blocks, describing deviations of average effects within a block from the global average. It is assumed to be distributed as $u \sim N(0, \sigma_u^2)$. Missing observations within any block and missing varieties between blocks lead to a reduced contribution to the estimates of variety effects.

Treating blocks as a random effect relies on interpreting blocks in this experiment as a sample of all possible blocks. Indeed, because the MM takes account of the variation in block effect it allows more reliable general conclusions to be drawn. The MM includes an estimate of the variability of block effect in the model, it may sometimes reduce statistical significance as a result, but it gives a more reliable indication of how robust the model will be if extended to other situations.

The MM may be extended so that the effects of variety are also treated as random terms; in this case uncertainty about block and variety information contribute to the shrinkage of estimates of effect.

EXAMPLE:

COMPLETE DATA				INCOMPLETE DATA		
Variety	Block	Yield		Variety	Block	Yield
Alpha	A	5.66		Alpha	B	8.26
Alpha	B	8.26		Alpha	C	7.40
Alpha	C	7.40		Alpha	D	7.23
Alpha	D	7.23		Alpha	A	8.04
Alpha	A	8.04		Alpha	B	3.61
Alpha	B	3.61		Alpha	C	9.92
Alpha	C	9.92		Alpha	D	11.03
Alpha	D	11.03		Beta	A	8.33
Beta	A	8.33		Beta	B	5.06
Beta	B	5.06		Beta	C	11.68
Beta	C	11.68		Beta	D	11.73
Beta	D	11.73		Beta	A	3.44
Beta	A	3.44		Beta	B	8.15
Beta	B	8.15		Beta	C	7.99
Beta	C	7.99		Beta	D	10.56
Beta	D	10.56		Gamma	A	8.91
Gamma	A	8.91		Gamma	B	8.22
Gamma	B	8.22		Gamma	C	11.81
Gamma	C	11.81		Gamma	D	9.52
Gamma	D	9.52		Gamma	A	10.25
Gamma	A	10.25		Gamma	B	8.85
Gamma	B	8.85		Gamma	C	9.64
Gamma	C	9.64		Gamma	D	9.94
Gamma	D	9.94		Delta	A	10.73
Delta	A	10.73		Delta	B	9.41
Delta	B	9.41				
Delta	C	10.11				
Delta	D	12.31				
Delta	A	11.41				
Delta	B	10.67				
Delta	C	8.07				
Delta	D	11.38				

Data simulation parameters.

Variety effects: Alpha,Beta,Gamma,Delta = 1,2,3,4. Block Effects: A,B,C,D = 1,2,3,4. Environmental error ~ *N(*4,2.5*)* =Ve.

Hence:

Genetic variance = Vg = variance (1,2,3,4)	=	1.6667
Phenotypic variance for this trial design (complete data) = Vp	=	Vg+Ve/8.
	=	1.6667+(2.5/8)
	=	1.9792
Theoretical h^2 for a variety from this experiment	=	Vg/Vp
	=	0.84

Statistical Analysis of Complete data.

ANOVA with the effect of variety as a fixed term – model (1).

Response:	Yield				
	Df	SumSq	MeanSq	F	Pr(>F)
Variety	3	39.435	13.145	3.0498	0.04493
Residuals	28	120.682	4.31		

The effect of variety is significant and heritability may be estimated since VarietyMS is an estimate of the population variance comprising 8Vg+Ve and the ResidualMS is the population estimate of Ve.

Thus 8Vg = 13.145-4.31 = 8.835 and Vg = 1.104.

Vp = Vg+Ve; in this case Vg is based on a sample of 8 plots of the variety (4 blocks x 2 replicates) so the appropriate value for Ve is 4.31/8 (the variance of the sampling distribution) so Vp = 1.643.

Heritability = 1.104/1.643 = 0.67.

ANOVA with the effect of variety and block as fixed terms – model (2).

Response:	Yield				
	Df	SumSq	MeanSq	F	Pr(>F)
Variety	3	39.435	13.145	3.8388	0.02176
Block	3	35.076	11.692	3.4144	0.03282
Residuals	25	85.607	3.424		

The inclusion of block effects allows more of the residual variation to be accounted for which makes the variety effect more significant than in the simpler model. Heritability rises to 0.74.

Mixed effect model with the effect of variety fixed and block as a random term.

Response:	Yield				
	Df	SumSq	MeanSq	F	Pr(>F)
Variety	3	39.435	13.145	3.8388	0.022

Estimates of the variance components from the model are:

Variety: (13.145-3.4244)/8 = 1.2151 = Vg

Residual: 3.4244 = Ve

Although the modelling method is inherently more sophisticated, when the data are complete the apportionment of variance is identical to the previous ANOVA and the values of F-ratio for variety and h^2 are unchanged.

Mixed effect model with the effect of variety random and block as a random term.

Estimates of the variance components from the model are:

Variety (Vg): 1.2151

Block (Vb): 1.0335

Residual (Ve): 3.4242

Giving an estimate of heritability of 0.74. These estimates of variance components are identical to those within the two-way ANOVA (for example; (1.2151 x 8) + 3.4242 = 13.145, etc) and it would be possible to construct an ANOVA table and perform an F-test. The MM, however, is more frequently used when data are incomplete. In this case it is not possible to estimate VarietyMS from the MM variance components since each variety may have been represented a different number of times. A significance test may still be constructed based on the change in deviance between the full MM and the MM with variety omitted; this difference being asymptotically distributed ~ χ_1^2 (See; Galwey 2006).

Likelihood ratio test.

Since H_0: $\sigma^2_\phi = 0$ which in ANOVA is tested in the upper tail of the F distribution then the probability from the inherently two-tailed χ^2_1 test must be halved to give the corresponding one-tailed estimate.

Estimates of change in deviance from REML in R are:

	Deviance	Change In deviance
Full model (block and variety as random terms):	137.3169	
Reduced model (block as random term, variety omitted):	140.7169	3.4
Reduced model (variety as random term, block omitted):	140.0748	2.76

Significance of variety effect = 0.5 x P(χ^2_1 >3.4) = 0.033

Significance of block effect = 0.5 x P(χ^2_1 >2.76) = 0.048

These values differ from the ANOVA estimates probably because the data-set is small and the distribution of the observed error terms departs from normality. However, the conclusion that both variety and block effects are significant (P=0.05) is unchanged.

Estimates of variety effects.

Ultimately the effect of variety on yield might be of interest and in this complete data set the estimate does not change provided there is more than one fixed term in the model. When two random terms are used the imprecision with which the variety and block effects have been made allows the model to attribute some of the differences in effect size to sampling variation and reduce their magnitude.

Estimates of variety effects (relative to overall trial mean) from Complete Data.

Variety	Known effect size (from simulation parameters)	Mean Yield	Effect (by difference)	Fixed model:Varietyfixed	Fixed model: Variety and Block fixed	MM: Variety fixed, Block random	MM: Variety and Block random
Alpha	-1.5	7.64	-1.40	-1.40	-1.40	-1.40	-1.03
Beta	-0.5	8.37	-0.67	-0.67	-0.67	-0.67	-0.50
Gamma	0.5	9.64	0.60	0.60	0.60	0.60	0.44
Delta	1.5	10.51	1.47	1.47	1.47	1.47	1.09

The estimates from ANOVA and the MM with variety fixed are identical; these equate to Best Linear Unbiased Estimates (BLUEs) of effects. The random effects model shrinks these estimates to reflect the uncertainty with which they are known, in the case of the complete balanced data the shrinkage factor is identical for each variety such that:

BLUP ≈ 0.74 x BLUE, note that the shrinkage factor is equivalent to the heritability calculated for the two-way ANOVA; the effect has been scaled down by the proportion of the total variance of variety means attributable to variety effect.

Statistical Analysis of incomplete data

ANOVA with the effect of variety as a fixed term.

Response:	Yield				
	Df	SumSq	MeanSq	F	Pr(>F)
Variety	3	15.73	5.243	1.0502	0.3911
Residuals	21	104.851	4.993		

The power of the experiment has been substantially reduced because the data are sparse and variety no longer appears significant.

Calculation of heritability is complicated because the amount of replication per variety is variable; a solution to this (Snedecor and Cochran 1973) is to use the harmonic mean of the number replications as a denominator in order to calculate sample variances. On this basis, from the table above varieties are replicated 4.48 times. Heritability is estimated as h^2 = 0.048.

ANOVA with the effect of variety and block as fixed terms.

	Df	SumSq	MeanSq	F	Pr(>F)
Block	3	29.964	9.988	2.7586	0.07228
Variety	3	25.446	8.482	2.3426	0.10732
Residuals	18	65.172	3.621		

Including the effect of blocks in the analysis allows more of the residual variance to be accounted and heritability rises to h^2 = 0.57. Note that neither the effect of blocks nor that of varieties is significant.

Mixed effect model with the effect of variety fixed and block as a random term.

Response:	Yield				
	Df	SumSq	MeanSq	F	Pr(>F)
Variety	3	22.244	7.4147	2.0456	0.143

Components of variance:

Variety	(7.4147-3.6247)/4.48=0.8459
Block (random)	1.61
Residual	3.6247

The mixed effects model takes account of the missing data by noting the uncertainty with which the block effects are known and so shrinking the block effects. This reduces the variance accounted by block and to some extent, variety. The consequence is that the estimate of heritability falls to $h^2 = 0.51$; however this may be a more reliable estimate. Despite the increased sophistication of the model the effect of variety on yield is not statistically significant in the incomplete data set.

Mixed effect model with the effect of variety random and block as a random term.

Estimates of the variance components from the model are:

Variety: 0.470

Block: 1.089

Residual: 3.894

With both block and variety terms random the average number of replications of variety is difficult to quantify. Galwey (2011) indicates that there is a continuing debate on the matter. However GENSTAT offers a potential solution since if stratum variances are requested the ratio between these (less residual variance) and the variance components is the denominator used. In this case the value is 6.34 which gives an estimate of heritability of 0.43. The estimates of variety and block effects have both been shrunk to account for the uncertainty introduced by missing data, consequently the amount of variance accounted by block and variety falls and residual variance rises.

The significance of variety and block effects may be estimated from the change in deviance (as above) and gives:

Significance of variety effect = $0.5 \times P(\chi_1^2 > 1.9) = 0.14$

Significance of block effect = $0.5 \times P(\chi_1^2 > 1.54) = 0.11$

The random effects model finds neither block nor variety effects significant, the p-value for variety from the LRT is similar in magnitude to that obtained from the F-test on the MM with variety as a fixed effect.

Ultimately the effect of variety on yield might be of interest and in this incomplete data it can be seen that the estimate of effect changes depending on the model used.

Estimates of variety effects (relative to overall trial mean) from incomplete data.

Variety	Known effect size (from simulation parameters)	MeanYield	Effect (by difference)	VarietyEffect:(Varietyfixed)	Variety Effect: (Variety and Block fixed)	Variety Effect: (Variety fixed, Block random)	Variety Effect: (Variety and Block random)
Alpha	-1.5	7.93	-0.86	-0.86	-1.90	-0.95	-0.36
Beta	-0.5	8.37	-0.42	-0.42	-1.33	-0.42	-0.15
Gamma	0.5	9.64	0.85	0.85	-0.05	0.85	0.47
Delta	1.5	10.07	1.28	1.28	1.64	2.19	0.47

The simplest fixed effect model can only estimate the effect of variety in terms of the difference between the data means taking no account of the very sparse data for Delta; this is clearly a poor estimate since Delta is only present in the lower yielding blocks. The fixed effect model including block effects is able to revise the estimate of effect for Delta upwards, however it notionally awards the full effect of the higher yielding blocks without adjustment for the uncertainty with which these or the available values for Delta are known. Thus the estimate of the effect of Delta is known with much less certainty than that for the other varieties since it is based on a small sample of the effects of Delta in a small sample of blocks.

Considering the mixed effects model; the simpler model (varieties fixed) recognises that variety Delta has only been tested in the lower yielding blocks, the estimate of its effect is higher than that obtained by difference or by the simplest fixed model. However, the magnitude of the variety effect is reduced relative to the two-term fixed model to take account of the incomplete sample taken of block effects. This is apparent in the effect sizes of Varieties Beta and Gamma which are essentially unchanged relative to the two-way ANOVA.

If both block and variety are treated as random the estimated effects are substantially reduced to account for both incomplete block and incomplete variety information. The effect size predicted for Delta has been shrunk such that it is identical to Gamma; this reflects the paucity of data for Delta. The effect sizes for Beta and Gamma are also shrunk relative to the estimates in the complete data to account for the missing block information.

The shrinkage factor between the BLUEs where variety is fixed and the BLUPs where both variety and block are random is proportional to the amount of data available; greatest shrinkage occurs with variety Delta which has the most incomplete data matrix. As indicated earlier, this shrinkage is also the specific heritability for the trait/variety combination in the experiment.

Variety	Variety Effect: (Variety fixed, Block random)	Variety Effect: (Variety and Block random)	Shrinkage factor (variety specific h^2)
Alpha	-0.95	-0.36	0.38
Beta	-0.42	-0.15	0.36
Gamma	0.85	0.47	0.55
Delta	2.19	0.47	0.21

Use of BLUPs and BLUEs

In essence the BLUEs are estimates of fixed effect from the model (be it ANOVA or MM). BLUPs are estimates of random effects inferred from a model in which the error structure of varieties and blocks have been inferred from the data. BLUPs are shrunk towards the mean effect which has the disadvantage of reducing the magnitude of the difference between the effects of varieties but this is almost always more than compensated for (see; Piepho, et al. 2008) by the reduction in the variance of these estimates(BLUPS and BLUES converge as

datasets become larger). Thus if an effect of variety is to be estimated from incomplete data the empirical BLUPs are generally a safer choice since they are known with greater accuracy.

MMs incorporating pedigree or kinship data.

In the example above varieties were considered as independent factors, each with an effect on yield which could be estimated. In that case, if varieties were treated as random terms then, essentially, the matrix Z_v was an identity matrix and the random effects of varieties were assumed to be distributed $u \sim N(0, \sigma_u^2)$ - uniform for all varieties.

However, if varieties are not independent of each other (i.e. there is a known kinship structure) then the error structure is more meaningfully distributed in proportion to this structure (the greater the coancestry the lower the proportion of genetic variation). In this case $Z_v$can be a variance/covariance matrix where the identity matrix is a special case in which varieties are independent. Thus the variance/covariance matrix serves to model the inter-relationships between factors and allocate σ_u^2 accordingly.

The model is:

$$y = X\beta + Zu + e$$

Where:

$$Var(u) = \sigma_g^2 K$$

K is a matrix of kinship coefficients and X is a matrix of fixed effects. If marker genotype is one of the fixed effects then this is the basis of association mapping using the MM (Yu *et al.* 2006) giving an estimate of association adjusted for kinship.

12.2 Appendix 2. Performance of regression models in the simulated cereal population with varying numbers of individuals, markers and loci.

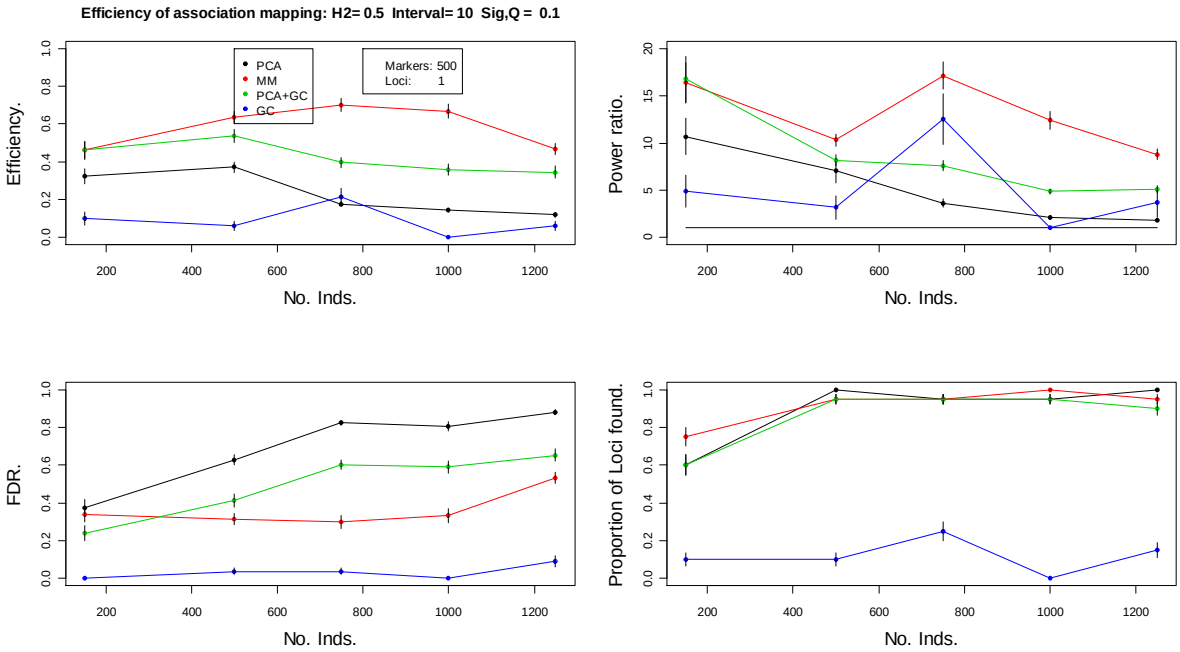


Figure 12-1. ***Power, FDR and Efficiency of regression models used to detect association with phenotype simulated in the simulated cereal data over number of randomly selected individuals in the panel. Loci account for 50% of phenotypic variation. Genotypes consisted of 500 randomly selected markers. Significance threshold was q=0.1. Detection is considered true within 10cM of the causative locus. Points are average of 100 simulations (error bars ± 1 s.d.).***

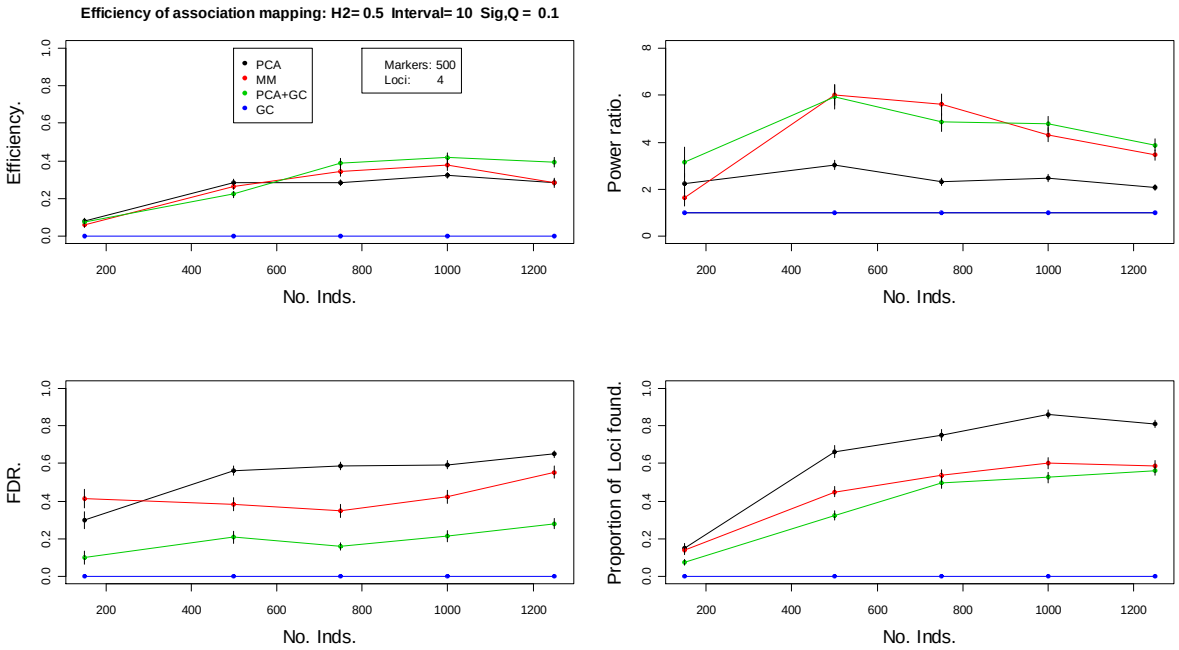


Figure 12-2. ***Power, FDR and Efficiency of regression models used to detect association with phenotype simulated in the simulated cereal data over number of randomly selected individuals in the panel. Loci account for 12.5% of phenotypic variation. Genotypes consisted of 500 randomly selected markers.***

Significance threshold was q=0.1. Detection is considered true within 10cM of the causative locus. Points are average of 100 simulations (error bars ± 1 s.d.).

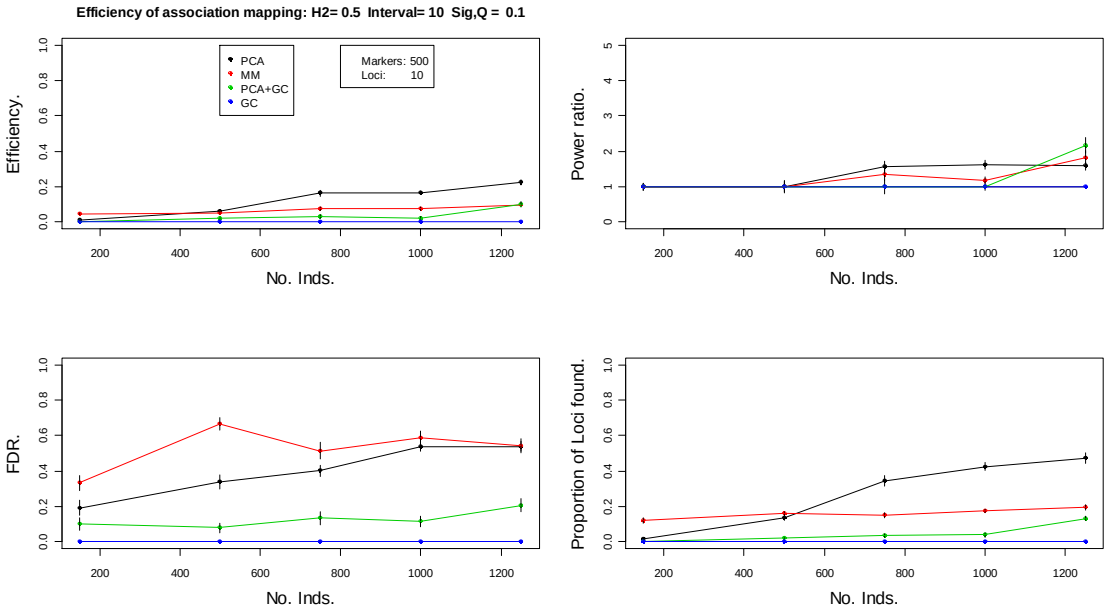


Figure 12-3. ***Power, FDR and Efficiency of regression models used to detect association with phenotype simulated in the simulated cereal data over number of randomly selected individuals in the panel. Loci account for 5% of phenotypic variation. Genotypes consisted of 500 randomly selected markers. Significance threshold was q=0.1. Detection is considered true within 10cM of the causative locus. Points are average of 100 simulations (error bars ± 1 s.d.).***

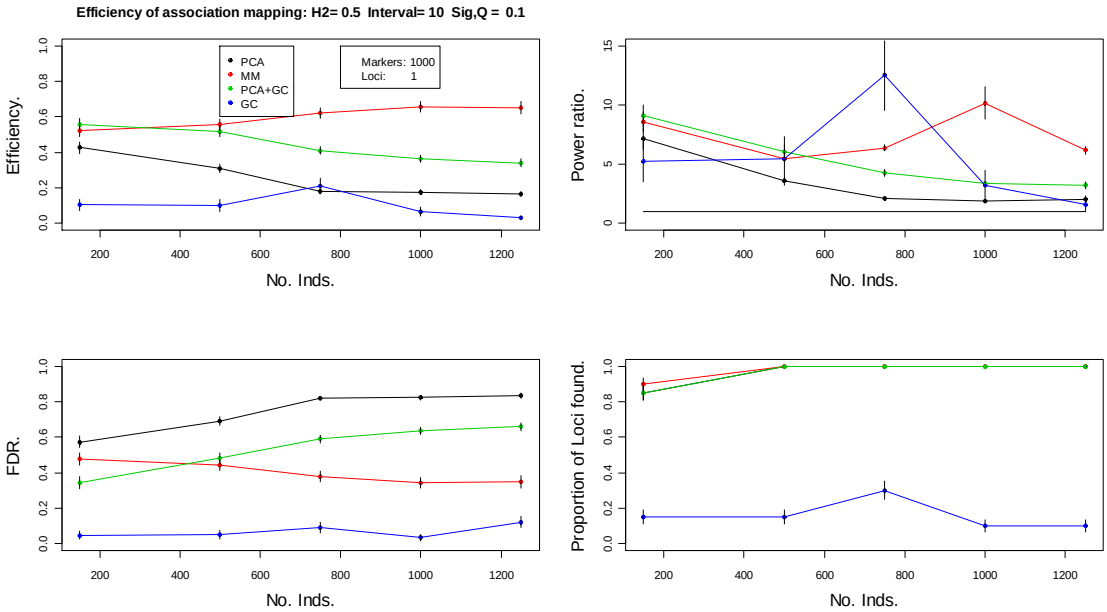


Figure 12-4. ***Power, FDR and Efficiency of regression models used to detect association with phenotype simulated in the simulated cereal data over number of randomly selected individuals in the panel. Loci account for 50% of phenotypic variation. Genotypes consisted of 1000 randomly selected markers. Significance threshold was q=0.1. Detection is considered true within 10cM of the causative locus. Points are average of 100 simulations (error bars ± 1 s.d.).***

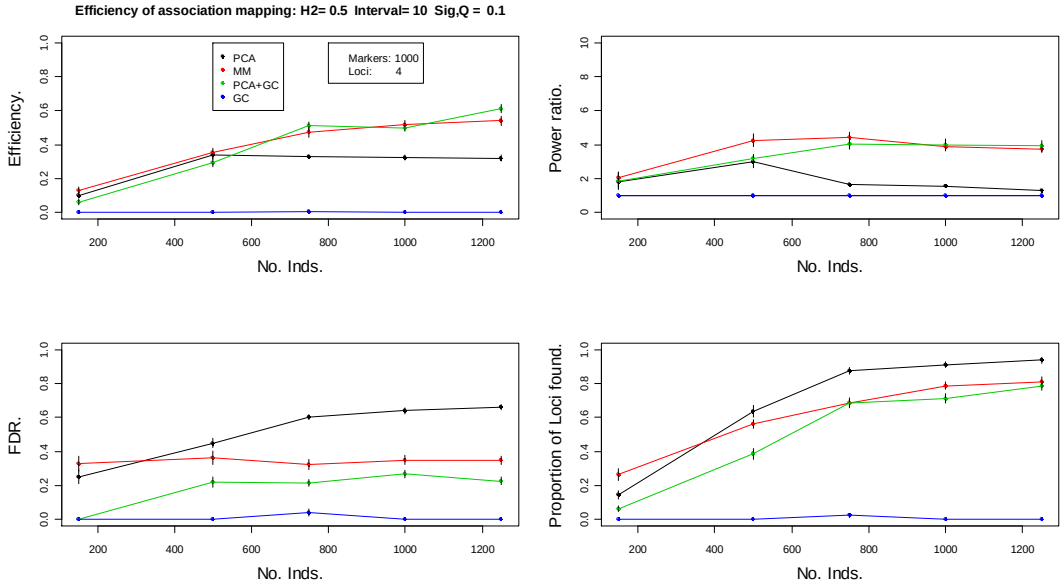


Figure 12-5. ***Power, FDR and Efficiency of regression models used to detect association with phenotype simulated in the simulated cereal data over number of randomly selected individuals in the panel. Loci account for 12.5% of phenotypic variation. Genotypes consisted of 1000 randomly selected markers. Significance threshold was q=0.1. Detection is considered true within 10cM of the causative locus. Points are average of 100 simulations (error bars ± 1 s.d.).***

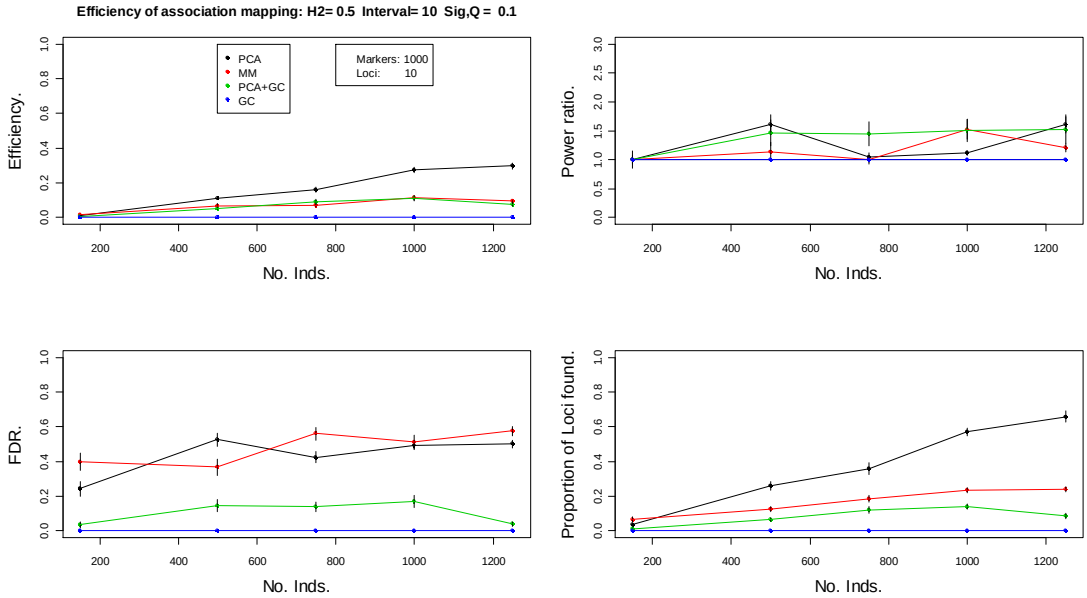


Figure 12-6. ***Power, FDR and Efficiency of regression models used to detect association with phenotype simulated in the simulated cereal data over number of randomly selected individuals in the panel. Loci account for 5% of phenotypic variation. Genotypes consisted of 1000 randomly selected markers. Significance threshold was q=0.1. Detection is considered true within 10cM of the causative locus. Points are average of 100 simulations (error bars ± 1 s.d.).***

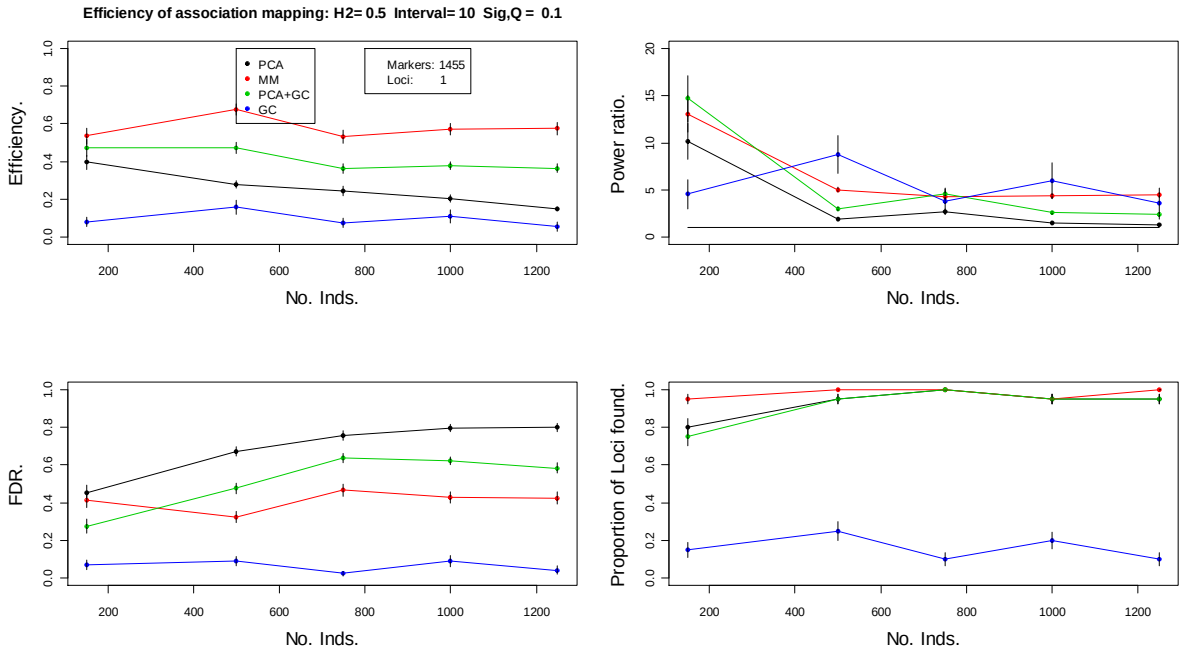


Figure 12-7. Power, FDR and Efficiency of regression models used to detect association with phenotype simulated in the simulated cereal data over number of randomly selected individuals in the panel. Loci account for 50% of phenotypic variation. Genotypes consisted of 1455 randomly selected markers. Significance threshold was q=0.1. Detection is considered true within 10cM of the causative locus. Points are average of 100 simulations (error bars ± 1 s.d.).

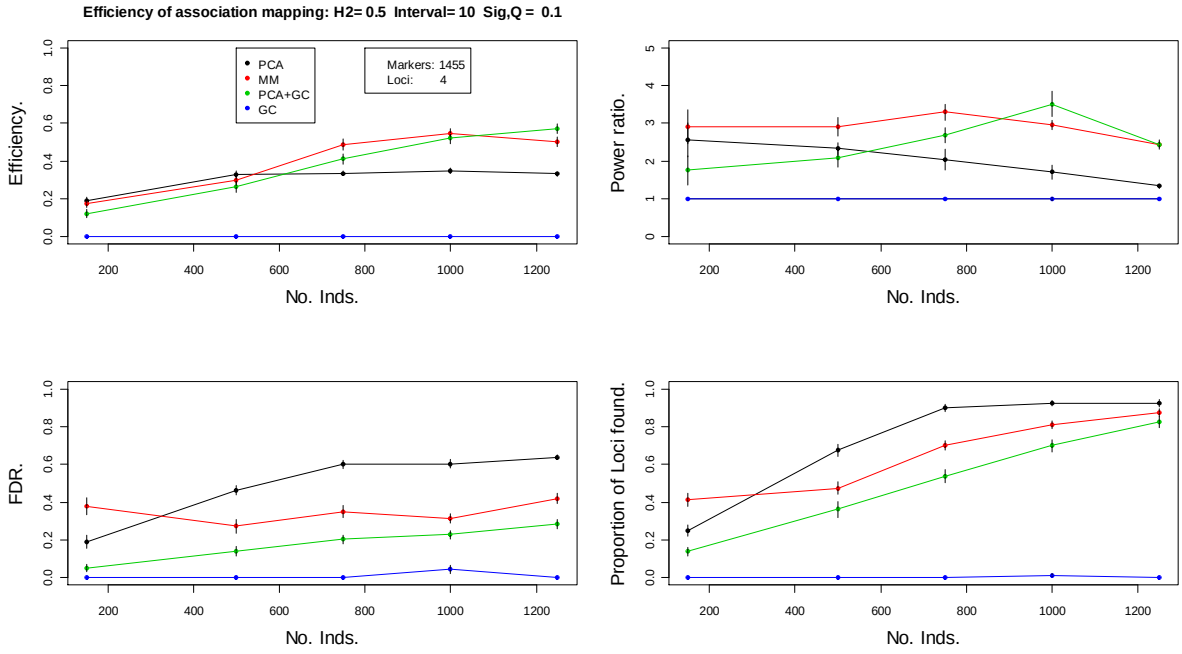


Figure 12-8. ***Power, FDR and Efficiency of regression models used to detect association with phenotype simulated in the simulated cereal data over number of randomly selected individuals in the panel. Loci account for 12.5% of phenotypic variation. Genotypes consisted of 1455 randomly selected markers. Significance threshold was q=0.1. Detection is considered true within 10cM of the causative locus. Points are average of 100 simulations (error bars ± 1 s.d.).***

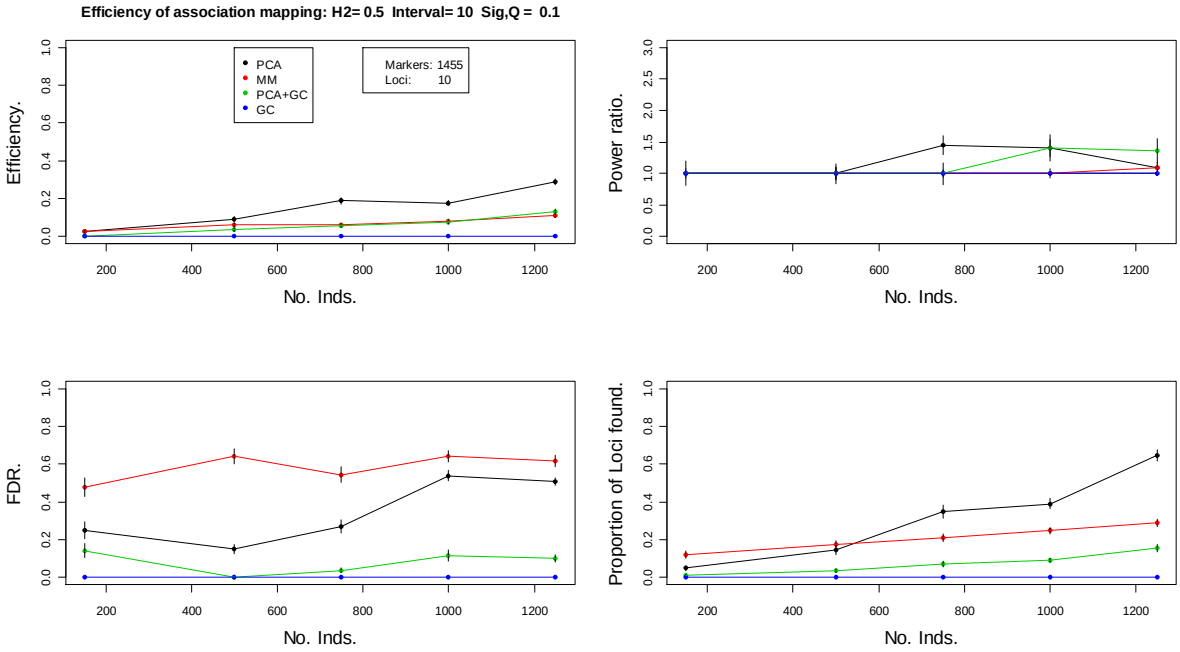


Figure 12-9. ***Power, FDR and Efficiency of regression models used to detect association with phenotype simulated in the simulated cereal data over number of randomly selected individuals in the panel. Loci account for 5% of phenotypic variation. Genotypes consisted of 1455 randomly selected markers. Significance threshold was q=0.1. Detection is considered true within 10cM of the causative locus. Points are average of 100 simulations (error bars ± 1 s.d.).***

12.3 Appendix 3. Barley varieties in the AGOUEB Panel.

AFP accession	Variety name	AFP accession	Variety name	AFP accession	Variety name
395	Plaisant	1382	Pacific	1764	Vortex
398	Halcyon	1384	Oleron	1766	Colston
456	Klaxon	1385	Becket	1768	Novello
560	Magie	1386	Flute	1769	Cocktail
578	Pipkin	1387	Chord	1770	Sebastian
597	Digger	1388	Mahogany	1773	Sw_Stella
610	Camargue	1389	Breeze	1775	Maypole
657	Marinka	1390	Magnolia	1777	Class
662	Torrent	1392	Bistro	1780	Brazil
674	Dandy	1394	Musette	1783	Louise
688	Blenheim	1396	Sevilla	1786	Sw_Hillary
721	Corniche	1397	Angela	1787	Sw_Sienna
725	Frolic	1398	Ethno	1788	Sw_Alison
733	Kira	1402	Heligan	1790	Connoisseur
750	Tyne	1403	Masquerade	1792	Nocturne
802	Prisma	1404	Tabetha	1793	Wombat
806	Puffin	1405	Molly	1794	Wigwam
816	Pastoral	1410	Century	1799	Kingston
820	Posaune	1411	Decanter	1804	Nord_96629_46_R2
827	Gaulois	1415	Linden	1808	Lambada
830	Melusine	1416	Thistle	1814	Maritem
836	Hart	1418	Mandolin	1820	Tallica
884	Gypsy	1419	Acapella	1823	Eden
888	Target	1420	Splash	1824	Mortimer
899	Clarine	1421	Z_91_103_21	1825	Camion
904	Alexis	1423	Horizon	1826	Godiva
910	Maud	1424	Spiral	1828	Aquarelle
946	Eagle	1426	Sabel	1835	Granta
947	Fighter	1427	Static	1836	Velvet
951	Manitou	1431	Pongo	1837	Macarena
954	Sprite	1432	Cecilia	1841	Mandolin
974	Chad	1435	Opal	1842	Amourette
994	Bronze	1436	Mariner	1847	Berillyum
1001	Karisma	1438	Antonia	1854	Nsl_99_5363
1011	Derkado	1441	Hurricane	1859	Skagen
1022	Dallas	1442	Wizard	1862	Carafe
1031	Chariot	1443	Steeple	1863	Athena
1052	Chestnut	1444	Zulu	1864	Toby
1055	Willow	1445	Arrow	1865	Doyen
1058	Swift	1448	Vanessa	1867	Drum

1059	Firefly	1450	Damas	1871	Rebecca
1071	Intro	1451	Vilna	1874	Indola
1091	Felicie	1453	Milano	1875	Feltwell
1101	Goldie	1456	Laurel	1880	Cpbt_B64
1107	Heron	1457	Rounder	1882	Cpbt_B66
1117	Gaelic	1463	Siberia	1884	Elmstead
1122	Finch	1466	Halifax	1885	Merode
1126	Epic	1467	Cpbt_B20_R2	1886	Amarena
1127	Linnet	1468	Artist	1888	Nord_96601_6
1133	Delibes	1469	Masai	1892	Faraday
1146	Cooper	1471	Leonie	1894	Coriolis
1147	Brewster	1474	Berwick	1899	Nectaria
1157	Juno	1480	Chime	1900	961374_R2
1160	Melanie	1488	Starlight	1902	12337_Zh_R2
1161	Duet	1490	Cindy	1903	Regalia
1162	Sunrise	1492	Potter	1906	Swub_01_41_R2
1165	Angora	1495	Foxtrot	1907	Calliope
1169	Rhythm	1496	Annabell	1908	Imogen
1170	Tempo	1498	Tavern	1910	Flagon
1171	Fanfare	1499	Chaser	1914	Rattle
1172	Madrigal	1500	Colada	1919	Thalia
1176	Digby	1501	Saloon	1920	Hermia
1179	Electron	1506	Whisper	1921	Mhhx011
1181	Tosca	1507	Cynthia	1923	Concept
1182	Gazelle	1510	Z_9175_R2	1924	Spectrum
1183	Emeraude	1511	Anvil	1933	Power
1184	Hanna	1514	Sonic	1937	Minstrel
1185	Amber	1527	Goldmine	1939	Westminster
1187	Cork	1528	Haka	1940	Oxbridge
1188	Optic	1529	Saffron	1946	Lithium
1191	Brahms	1530	Harland	1947	Cpbt_B67
1194	Canasta	1531	Campion	1949	Lp1124_8_98
1196	Reggae	1532	Chamomile	1950	Sw_Scania
1201	Jive	1534	Sarah	1951	Sw_Macsena
1202	Glen	1535	Carola	1955	Macaw
1203	Riviera	1536	Weaver	1957	Toucan
1204	Regina	1540	Violet	1961	Ursa
1206	Tudor	1541	Montage	1963	Waggon
1208	Druid	1544	Cwb_97_6_R2	1968	Wicket
1209	Punch	1547	Sumo	1974	Timori
1210	Tokyo	1552	Barcelona	1980	Charleston
1211	Arctic	1554	Thrift	1981	Houston
1212	Medoc	1555	Akita	1982	Celebrity

1213	Jerez	1556	Dew	1983	Cypress
1214	Muscat	1563	Alliot	1984	Cinnamon
1215	Pacific	1565	Astoria	1985	Dolphin
1219	Mystique	1566	Prestige	1987	Malwinta
1220	Ici_85161_4427	1567	Protege	1989	Archimedes
1221	Prelude	1574	Cellar	1990	Celsius
1224	Vanilla	1575	Pewter	1991	Fahrenheit
1225	Tamaris	1576	Brise	1996	Mosaic
1228	Gleam	1579	Viskosa	2002	Sw_Norma
1230	Crown	1580	Neruda	2003	Colibri
1232	Onyx	1581	Nsl_95_1257	2004	Monalisa
1236	Laird	1583	Nsl_97_4552	2007	Ac_97_H2406_10
1237	Chieftain	1584	Nsl_97_4579	2017	Silverstone
1238	Tankard	1589	Agenda	2021	Surtees
1239	Pitcher	1594	Fontana	2024	Sw_2808
1240	Trinity	1595	Avec	2025	Corsica
1242	Wren	1597	Alabama	2030	Appaloosa
1244	Ragtime	1599	Pippa	2031	Hydra
1245	Primera	1604	Diamond	2034	Rummy
1246	Clarity	1605	Sapphire	2035	Skittle
1248	Mikado	1606	Carat	2036	Poker
1250	Melitta	1608	Tucker	2039	Cribbage
1253	Tardus	1610	Tipster	2040	Silicon
1255	Polygena	1617	Chicane	2042	Kassima
1256	Trinidad	1619	Moonshine	2044	995964
1259	Kite	1621	Vesuvius	2047	Beatrix
1260	Falcon	1623	Godiva	2048	Aluminium
1261	Jet	1624	Milena	2050	Tucson
1262	Volley	1625	Aquarelle	2052	Turnberry
1263	Rifle	1626	Pict	2053	Putney
1264	Cobalt	1627	Perth	2054	Centurion
1267	Credo	1628	Scylla	2055	Paramount
1271	Glint	1630	Caption	2058	Cassata
1276	Antigua	1642	Tropic	2065	Retriever
1281	Emilia	1643	Spinner	2066	Marado
1282	Madison	1651	Sacha	2067	Selection
1283	Pilot	1654	Braemar	2071	5593_Bh2
1285	Spice	1659	Vegas	2073	Accrue
1287	Portrait	1663	Nsl_98_5065	2075	Sw_165
1290	Crusader	1666	Adonis	2076	Blythe
1291	Quartet	1668	Spire	2077	Pelican
1293	Landlord	1669	Spike	2078	Cebeco_02215_05
1294	Draught	1673	Celebra	2088	Cedar
1295	Host	1676	Harriot	2090	Mh_Bi_8

1297	Hopper	1679	Widre	2091	Mh_94_Cf_13_1_2
1298	Toddy	1680	Weitor	2092	Tartan
1301	Rainbow	1681	Wikingett	2096	Scandium
1303	Charm	1682	Heather	2098	Monika
1307	Henni	1687	Anais	2102	Acrobat
1311	Torup	1688	Tabora	2103	Token
1312	Renata	1693	Jessica	2104	Calico
1313	Trosa	1697	Sunbeam	2105	Vivendi
1316	Crescendo	1698	Clara	2106	Gundel
1317	Ravel	1700	Cannock	2110	Shakira
1318	Pearl	1701	Nsl_97_6002_R2	2111	Anaconda
1320	Jewel	1702	Saigon	2112	Cpbt_B76
1321	Peridot	1705	Kestrel	2119	Publican
1324	Honey	1706	Swallow	2121	Quench
1325	Lark	1708	Sombrero	2122	Taphouse
1326	Baton	1709	Eden	2123	Penthouse
1327	Toffee	1715	Pedigree	2130	Prague
1330	Turine	1717	Sequel	2134	Cpbt_B78
1334	Spirit	1721	Antelope	2135	Cpbt_B79
1335	Vertige	1722	Calcutta	2138	Nsl_03_7309
1339	Winner	1724	Cathay	2145	410_3e
1344	Ayana	1726	Outlook	2147	5353_Dh1
1345	Candy	1727	Lomerit	2160	Ac_99_077_2
1346	Askanova	1728	Mead	2161	Ac_997_077_13
1349	Livet	1730	Parasol	2171	Asb_04_18
1350	Ricarda	1731	Cellina	2173	Asb_04_20
1353	Dray	1733	Sw_Farrier	2180	Sb02146
1354	Chalice	1736	Boreale	2187	Nord_03_2408
1356	Ferment	1741	Rakaia	2188	Csbc_5466_27
1357	Extract	1744	Nimbus	2201	Nsl_04_4622
1361	Scarlett	1745	Pelican	2203	Nsl_04_4683
1363	Ardila	1746	Catalina	2206	Nfc_404_64
1364	Madras	1748	Kirsty	2209	Nfc_404_80
1375	Fractal	1755	Troon	2212	Nfc_404_62
1376	Spey	1758	Rangoon	2213	Cpbt_B80
1377	Chintz	1761	Global		
1381	Goldrush	1762	Campala		

12.4 Appendix 4. List of wheat varieties included in the final association panel. The reference number is that given by the testing authority when the breeder made an application for registration.

Ref	Approved Name	Ref	Approved Name	Ref	Approved Name	Ref	Approved Name	Ref	Approved Name
4	BOUQUET	321	NORMAN	858	RIALTO	1147	NAPIER	1395	SENATOR
18	M.RANGER	324	MITHRAS	867	PROPHET	1153	GENGHIS	1411	DICKSON
33	FLINOR	325	RAPIER	878	BEAUFORT	1156	BRANDO	1414	VECTOR
66	M.HUNTSMAN	350	PAGEANT	881	ENCORE	1160	COCKPIT	1426	ISTABRAQ
103	M.FREEMAN	364	LONGBOW	882	CONSORT	1200	OPTION	1427	SOCRATES
109	MEGA	370	FENMAN	886	TURPIN	1202	OXBOW	1440	STEADFAST
111	ATOU_A	440	GALAHAD	931	CAXTON	1205	POTENT	1443	HERITAGE
170	GAMIN	489	BROCK	932	REAPER	1211	ODYSSEY	1447	BENTLEY
173	FLANDERS	533	MERCIA	933	RALEIGH	1213	EXSEPT	1449	WELFORD
178	KINSMAN	591	HORNET	939	CHARGER	1220	DEBEN	1462	AMBROSIA
179	HOBBIT	607	APOLLO	943	GALATEA	1223	CANTERBURY	1468	DEFENDER
196	SPORTSMAN	628	RIBAND	946	CHIANTI	1224	CYBER	1477	ATLANTA
201	ARMADA	670	APOSTLE	954	CROFTER	1227	MARLIN	1482	GLASGOW
227	VALMY	671	PASTICHE	967	ABBOT	1244	ASHBY	1502	EARL
228	RIVOLI	692	BEAVER	970	IMPALA	1257	RICHMOND	1512	EXETER
230	HUSTLER	694	HAVEN	973	MADRIGAL	1258	CHATSWORTH	1541	PIRANHA
231	BRIGAND	724	EKLA	978	HARRIER	1262	MACRO	1542	DIRECTOR
233	MARDLER	728	TALON	980	MALACCA	1263	CLICK	1545	ZEBEDEE
236	MARKSMAN	735	TARA	983	EQUINOX	1266	ACCESS	1546	GATSBY
241	WAGGONER	736	HEREWARD	1029	BLAZE	1278	PHLEBAS	1549	FASTNET
243	AQUILA	759	ADMIRAL	1030	CANTATA	1281	XI19	1550	DOVER
260	ANVIL	775	ESTICA	1031	FALSTAFF	1282	SOLSTICE	1561	HYPERION
262	STUART	776	WASP	1033	SAVANNAH	1286	STORM	1564	ALCHEMY
265	COPAIN	787	TORFRIDA	1035	MAVERICK	1326	TELLUS	1577	HURLEY
270	WIZARD	800	GENESIS	1039	CHAUCER	1327	BRUNEL	1599	BATTALION
271	IONA	808	SPARK	1040	WESTON	1330	ROBIGUS	1621	GULLIVER
273	VILLEIN	810	ZODIAC	1047	KRAKATOA	1335	SCORPION	1625	DEACON
274	BOUNTY	817	HUSSAR	1048	WELLINGTON	1336	WARLOCK	1644	TIMBER
276	VIRTUE	818	BRIGADIER	1070	CLAIRE	1343	CARLTON	1649	SAHARA
287	AVALON	828	HUNTER	1092	SHAMROCK	1353	GOODWILL	1652	HUMBER
295	PRINCE	833	CADENZA	1100	BUCHAN	1376	EINSTEIN	1658	OAKLEY
310	BARON	845	BUSTER	1111	AARDVARK	1380	CARINE		
311	ABELE	847	FLAME	1126	ECLIPSE	1388	CORDIALE		
314	PACER	856	LYNX	1137	BROILER	1392	QUEST		

12.5 Appendix 5. Protocol For Modified Tanskley's DNA Microprep

THIS METHOD USES A NUMBER OF HAZARDOUS CHEMICALS. THESE ARE;

EDTA

Risk Phrases associated with this product

Irritating to the eyes, respiratory system and to the skin. Harmful if inhaled.

Safety phrases associated with this product

In case of contact with the eyes, rinse immediately with plenty of water and seek medical advice.

Wear suitable protective clothing

Likelihood of contamination of the operator or anyone else

Minimal as dilute solutions used.

Chloroform Isoamyl Alcohol (24;1)

Risk Phrases associated with this product

Harmful: danger of serious damage to health by prolonged exposure through inhalation or if swallowed. Toxic by inhalation.

Safety phrases associated with this product

In case of contact with the eyes, rinse immediately with plenty of water and seek medical advice

Take off immediately all contaminated clothing

After contact with skin, wash immediately with plenty of water

Wear suitable protective clothing.

Work in a fume hood.

Likelihood of contamination of the operator or anyone else

Minimal provided adequate safety procedures followed.

3. Isopropanol

Risk Phrases associated with this product

Highly flammable

Safety phrases associated with this product

Keep in a well ventilated place. Avoid sources of ignition.

Likelihood of contamination of the operator or anyone else

Minimal provided adequate safety procedures followed. Very small quantities are in use.

4. Tris

Risk Phrases associated with this product

Irritating to the eyes, respiratory system and skin

Safety phrases associated with this product

In case of contact with the eyes, rinse immediately with plenty of water and seek medical advice. Wear suitable protective clothing.

Likelihood of contamination of the operator or anyone else

Minimal provided adequate safety procedures followed.

5. Sorbitol

Risk Phrases associated with this product

May be harmful by inhalation, ingestion or skin adsorption.

Safety phrases associated with this product

Wear suitable protective clothing.

Likelihood of contamination of the operator or anyone else

Minimal provided adequate safety procedures followed as very small quantities are in use..

6. Sarkosyl

Risk Phrases associated with this product

May be harmful by inhalation, ingestion or skin adsorption.

Safety phrases associated with this product

Wear suitable protective clothing.

Likelihood of contamination of the operator or anyone else

Minimal provided adequate safety procedures followed as very small quantities are in use.

CTAB

Risk Phrases associated with this product

Corrosive; causes burns; harmful by inhalation, in contact with skin and if swallowed.

Safety phrases associated with this product

In case of contact with eyes, rinse immediately with plenty of water and seek medical advice.

Wear suitable protective clothing.

Likelihood of contamination of the operator or anyone else

Minimal provided adequate safety procedures followed as very small quantities are in use.

Sodium Bisulphite

Risk Phrases associated with this product

May be harmful by inhalation, ingestion or skin adsorption.

Safety phrases associated with this product

Avoid contact with skin and eyes. Wear suitable protective clothing.

Likelihood of contamination of the operator or anyone else

Minimal provided adequate safety procedures followed as very small quantities are in use.

Reagents Required

1M Sorbitol

5%Sarkosyl

DNA Extraction Buffer; *0.35M Sorbitol, 0.1M Tris, 5mM EDTA.*

Nuclei Lysis Buffer; *0.2M Tris, 0.5M EDTA, 2M NaCl, 2% CTAB.*

Chloroform : Isoamylalcohol, (24;1)

Isopropanol

70% Ethanol

Prepare Buffers and Autoclave

For all buffers use RO water

DNA Extraction Buffer; *0.35M Sorbitol, 0.1M Tris, 5mM EDTA*

Chemical	Stock	Quantity added for 250ml	Final concentration
Sorbitol	powder	15.9g	0.35M
Tris	powder	3.0g	0.1M
EDTA	0.5M solution	2.5ml	5mM
Water		Make up to 250ml	

Nuclei Lysis Buffer; *0.2M Tris, 0.5M EDTA, 2M NaCl, 2% CTAB.*

Chemical	Stock	Quantity added for 250ml	Final concentration
Tris	powder	6.1	0.2M
EDTA	0.5M solution	25ml	50mM
NaCl	powder	29.2g	2.0M
CTAB	powder	5g	2%
water		Make up to 250ml	

Sarkosyl; 5%

Chemical	Stock	Quantity added for 250ml	Final concentration
Sarkosyl	powder	12.5g	5%
water		250ml	

2. Make Microprep Buffer up immediately prior to use;

2.5x DNA Extraction Buffer (for 1 x 96 well block = 25 mls)

2.5x Nuclei Lysis Buffer (25 mls)

1x 5% Sarkosyl (10 mls)

0.2g Sodium Bisulfite

60µl RNase solution from qiagen

3. put balls into each well of a 96 well plate

4. harvest 50 - 100 mg young leaf tissue. If not using immediately, put plate without caps but with clear lid into the -80℃ freezer

5. remove balls from any empty wells

6. add 500µl buffer

7. grind using the mixer mill for 30s at 30hz. Turn plates and repeat – make sure the caps are on well. CAUTIONA. The mixer mill is very noisy. Leave the room when it is running. Make sure the plates are fixed securely into the clamps and that the safety cover is closed.

8. incubate at 65℃ for 30 – 120 minutes, shaking halfway through. Make sure the clear plastic lid is on the plate and place a rack on top of the lid as the tube lids may fly off in the oven.

9. after incubation cool the blocks for 15mins in the fridge

10. In the fume hood, add 300µl chloroform : isoamylalcohol. Mix by inverting the plate; make sure the caps are on tight, and there clear plastic lid is on too. Make sure that your gloves don't become contaminated with chloroform – change them if they do. Put all waste into the mixed hazardous bag in the fume hood.

11. Centrifuge the plates at 1500g for 5 mins.

12. pipette the aqueous phase off into a new 96 well plate. Use a 200µl pipette, work slowly and don't disturb the interface between the two phases. Recover the balls and wash in HCl. Decant the chloroform waste into the mixed hazardous waste bottle and place contaminated plastics into the solid hazardous waste bag for disposal.

13. At 340µl cold isopropanol to each well, and mix gently by inversion until the DNA is visable as a globule.

14. Centrifuge at 6000g for 5 mins.

15. pipette off the isopropanol. Add 500µl 70% ethanol and vortex briefly to lift the pellet from the bottom of the tube.

16. Centrifuge at 6000g for 5 mins; remove ethanol carefully with a pipette

17. repeat ethanol wash step

18. centrifuge for 1 min at 1000g to bring remainder of ethanol to bottem of tube.

19. dry at 65℃ for 10 min – don’t over dry; keep checking.

20. resuspend pellet in 100µl TE. Incubate at 65°C if necessary for 5 mins.

12.6 Appendix 6. Histograms of barley phenotype scores.

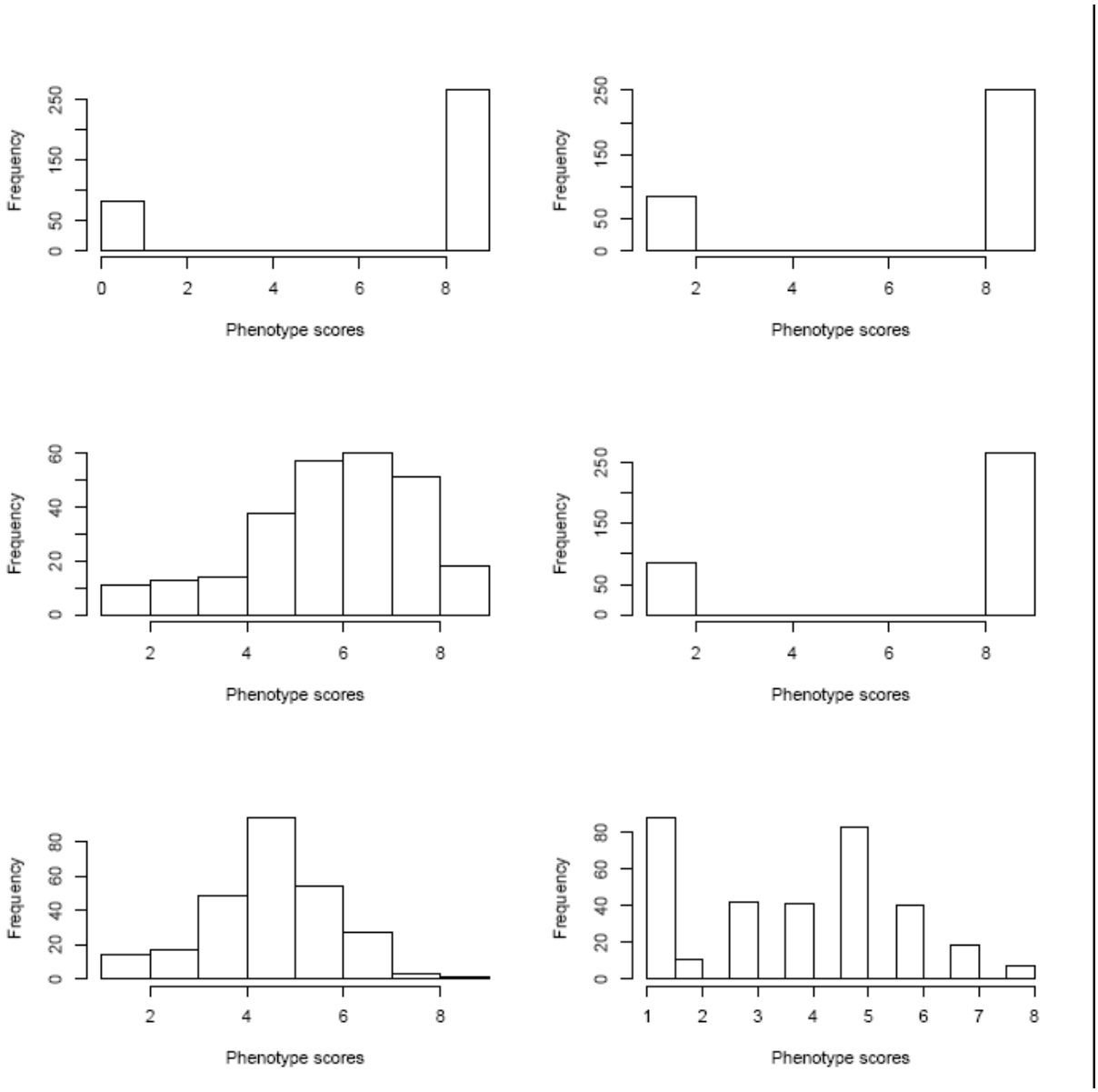


Figure 12-10. Distribution of phenotype scores for (left to right, top to bottom); Anthocyanin competence, Anthocyanin colouration of auricles, Intensity of anthocyanin colouration of auricles, Anthocyanin colouration of awn tips, Intensity of anthocyanin colouration of awn tips, Anthocyanin colouration of grain lemma nerves.

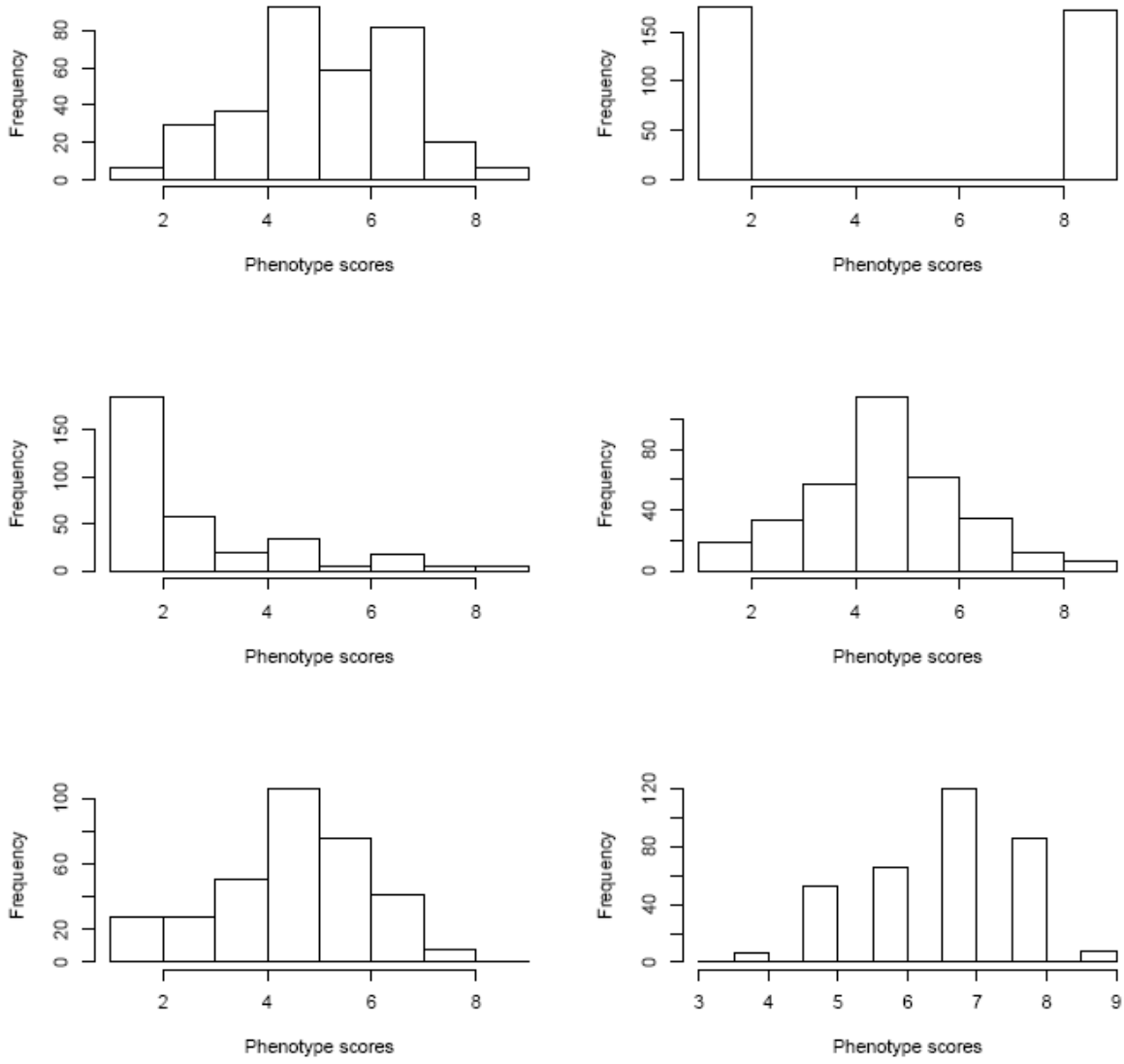


Figure 12-11. Distribution of phenotype scores for (left to right, top to bottom); Plant growth habit, Hairiness of the lower leaf sheaths, Frequency of plants with re-curved leaves, Time of ear emergence, Ear glaucosity, Glaucosity of the flag leaf sheath.

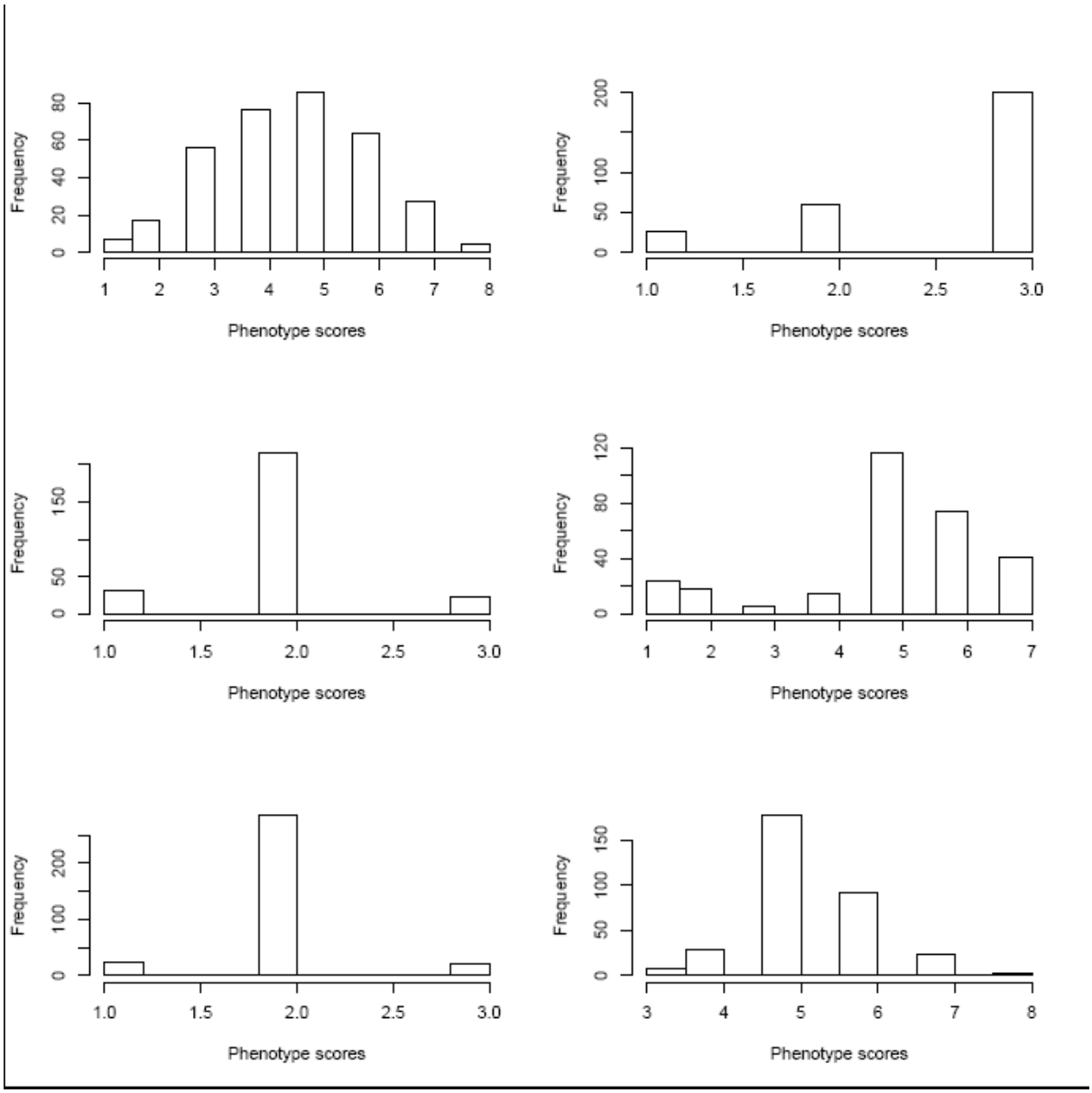


Figure 12-12. Distribution of phenotype scores for (left to right, top to bottom); Ear attitude, Sterile spikelet attitude, Shape of the tip of the sterile spikelet, Development of the sterile spikelet, Length of glume c.f. grain, Ear length.

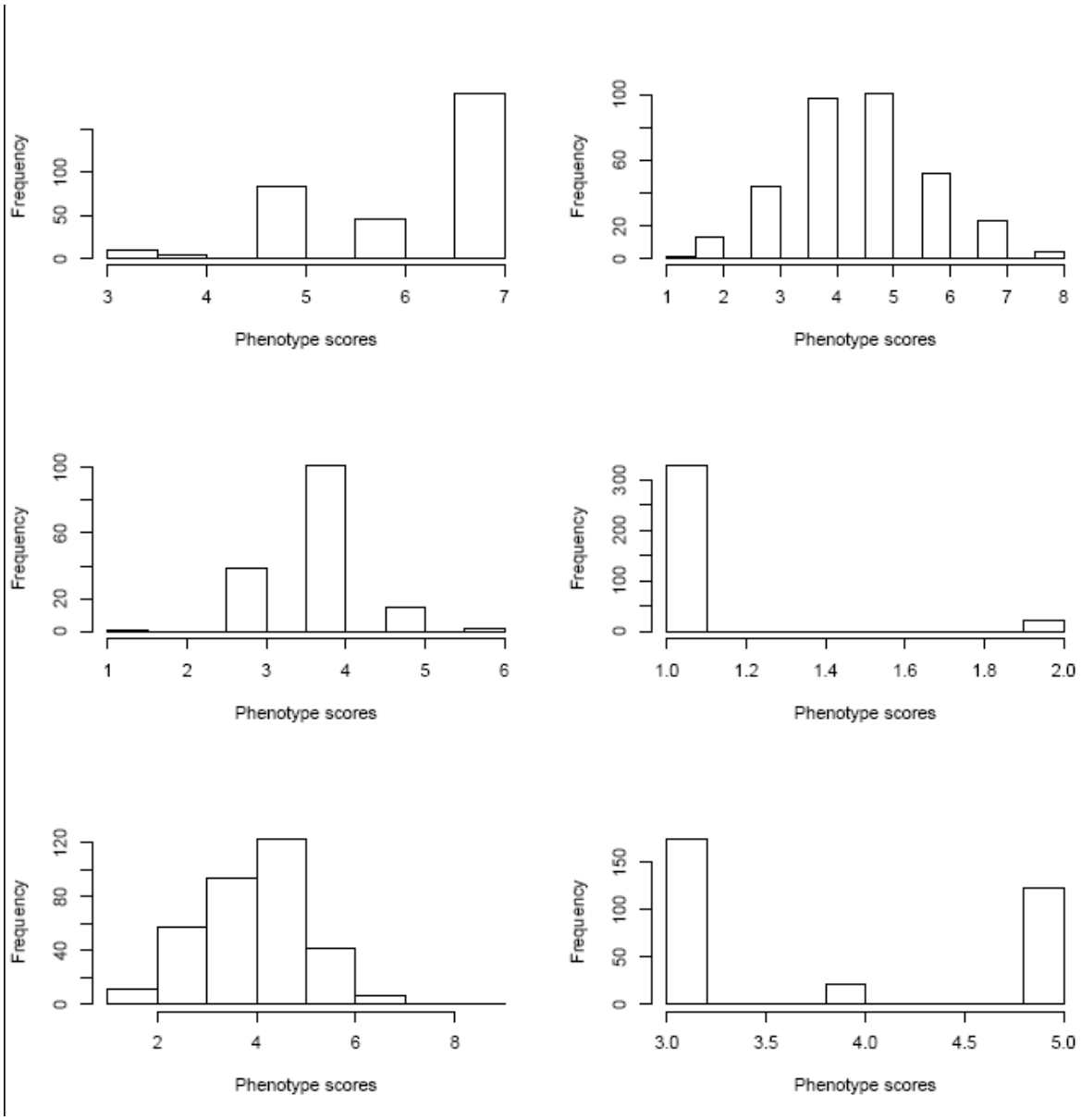


Figure 12-13. Distribution of phenotype scores for (left to right, top to bottom); Awn length, Plant length, Collar type, Row number (6/2), Ear density, Ear shape.

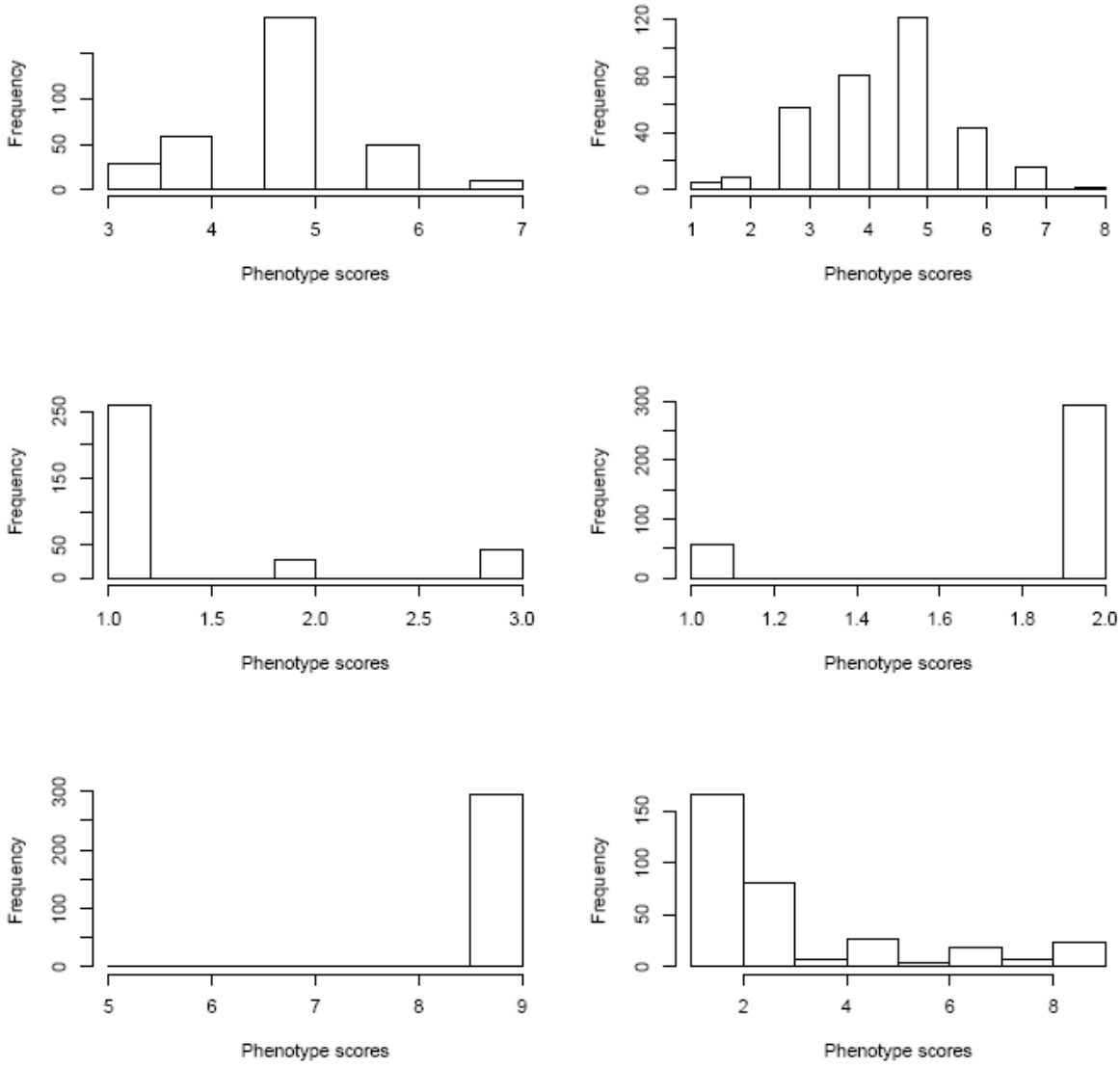


Figure 12-14. Distribution of phenotype scores for (left to right, top to bottom); Length of first rachis segment, Curvature of first rachis segment, Colour of the aleurone layer, Rachilla hair type, Spiculation of awn margins, speculation of inner lateral nerves.

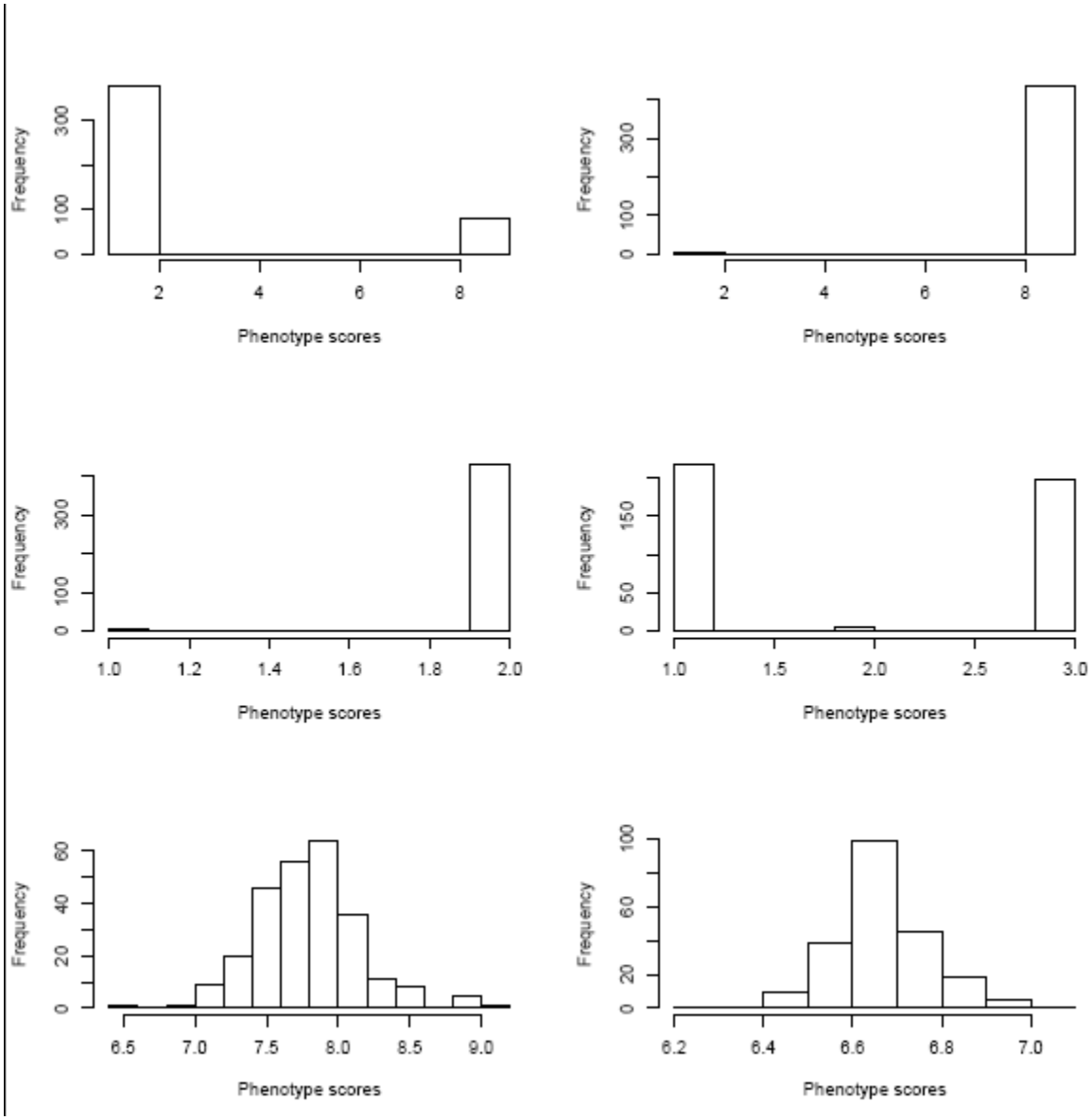


Figure 12-15. Distribution of phenotype scores for (left to right, top to bottom); Hairs in the ventral furrow, Adherence of grain husk, Disposition of lodicules, Seasonal type (W/Sp), Yield of winter sown varieties, Yield of spring sown varieties.

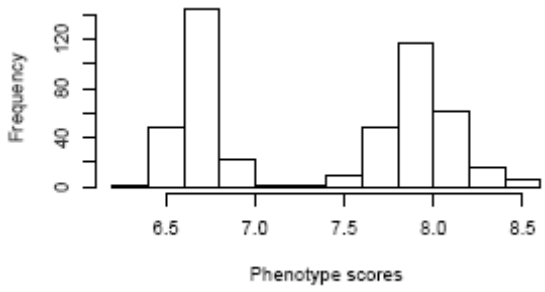


Figure 12-16. Distribution of phenotype scores for; Yield of winter and spring sown varieties together.

12.7 Appendix 7. Association plots for barley phenotypes. Association estimated without correction for structure.

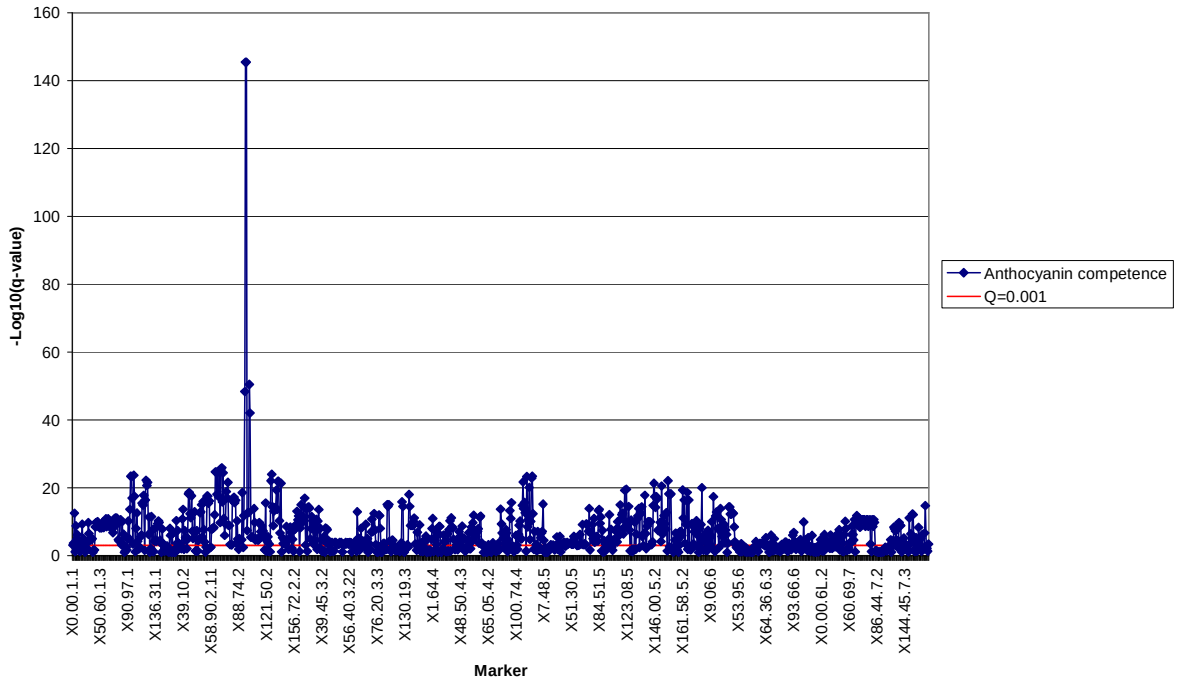


Figure 12-17. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

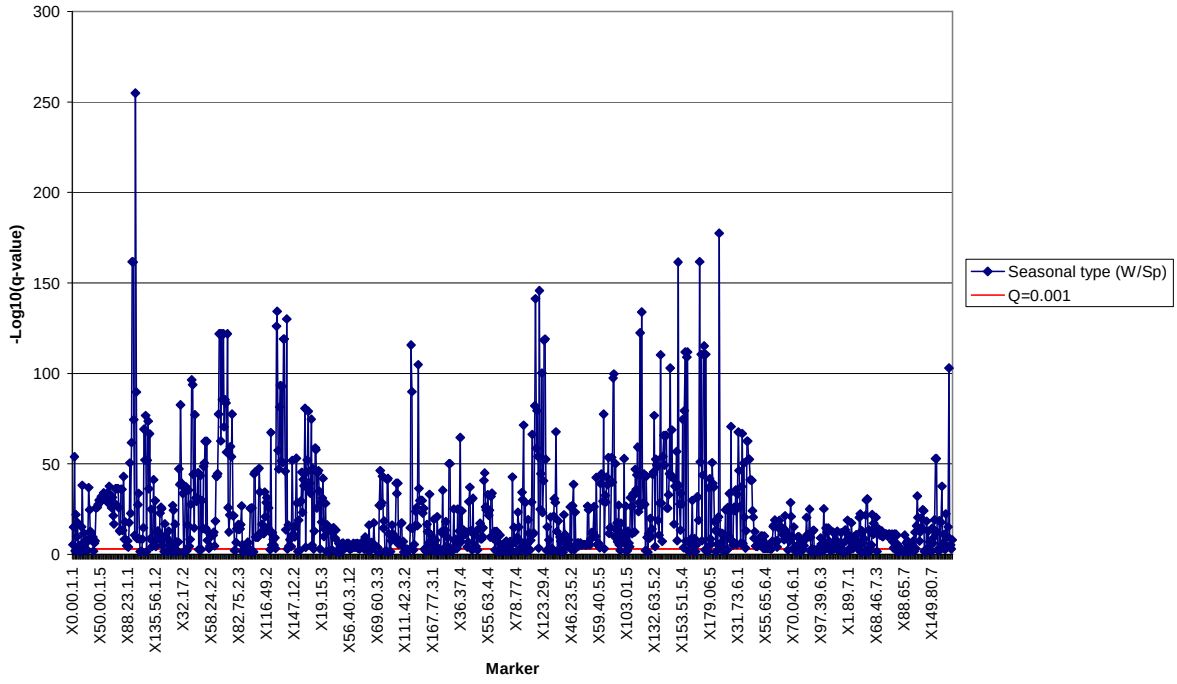


Figure 12-18. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

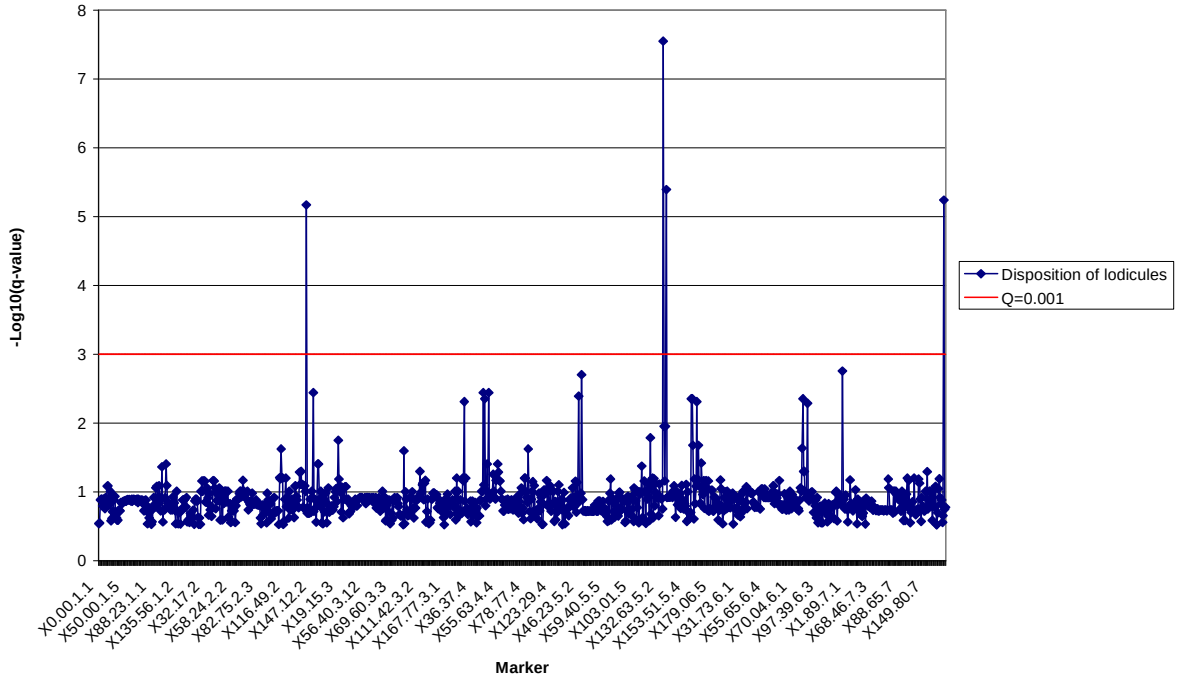


Figure 12-19. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

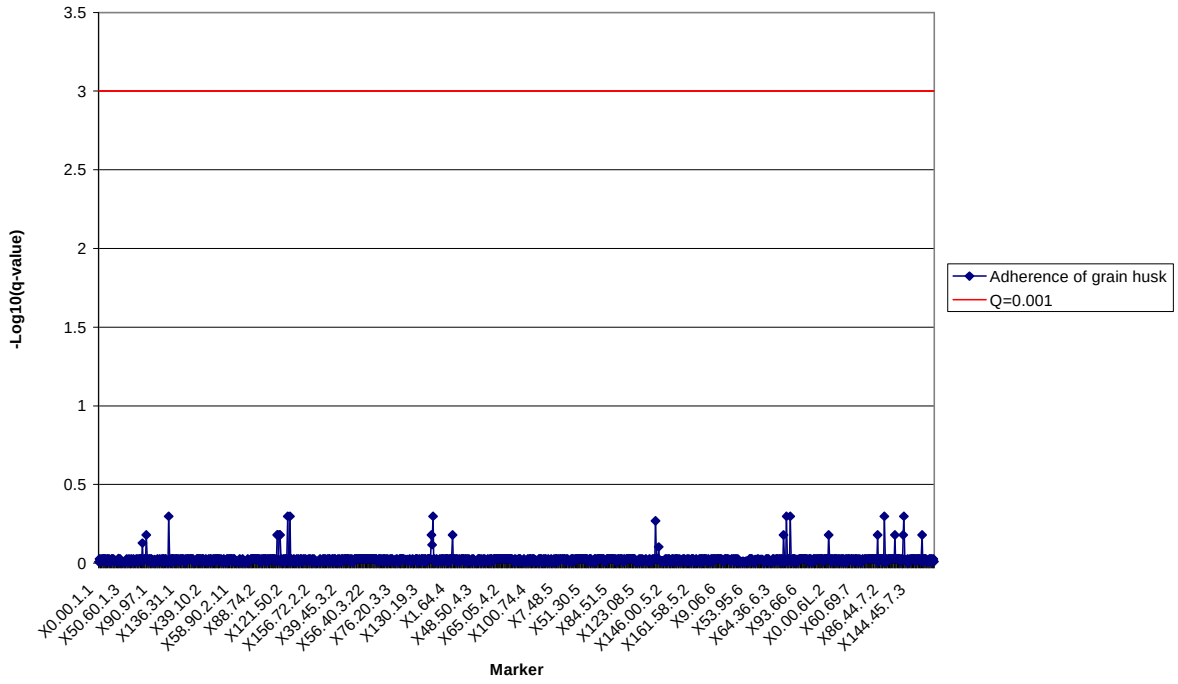


Figure 12-20. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

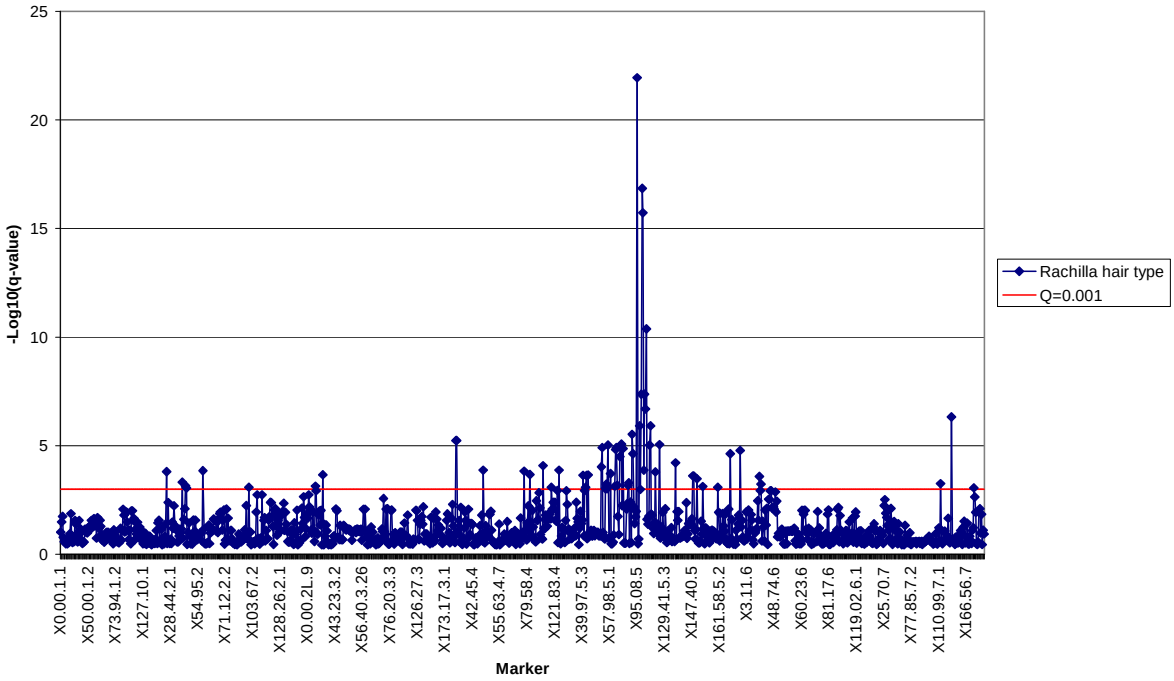


Figure 12-21. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

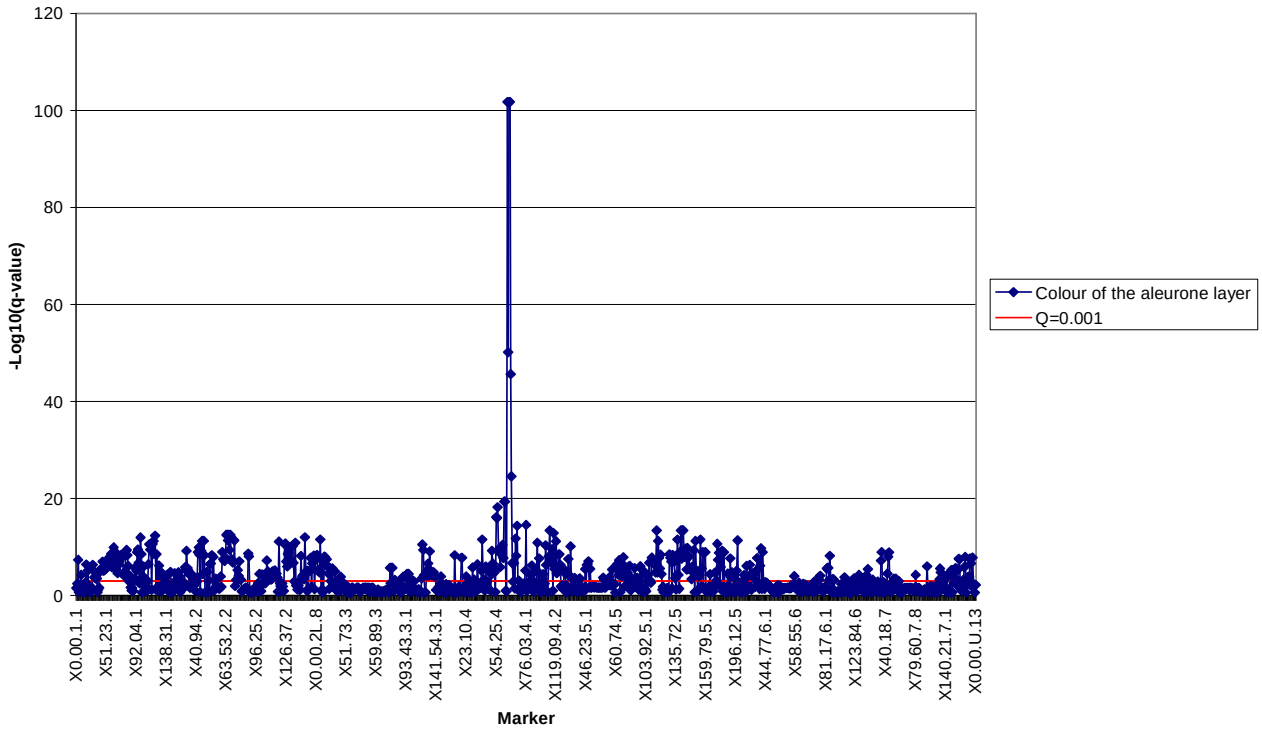


Figure 12-22. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

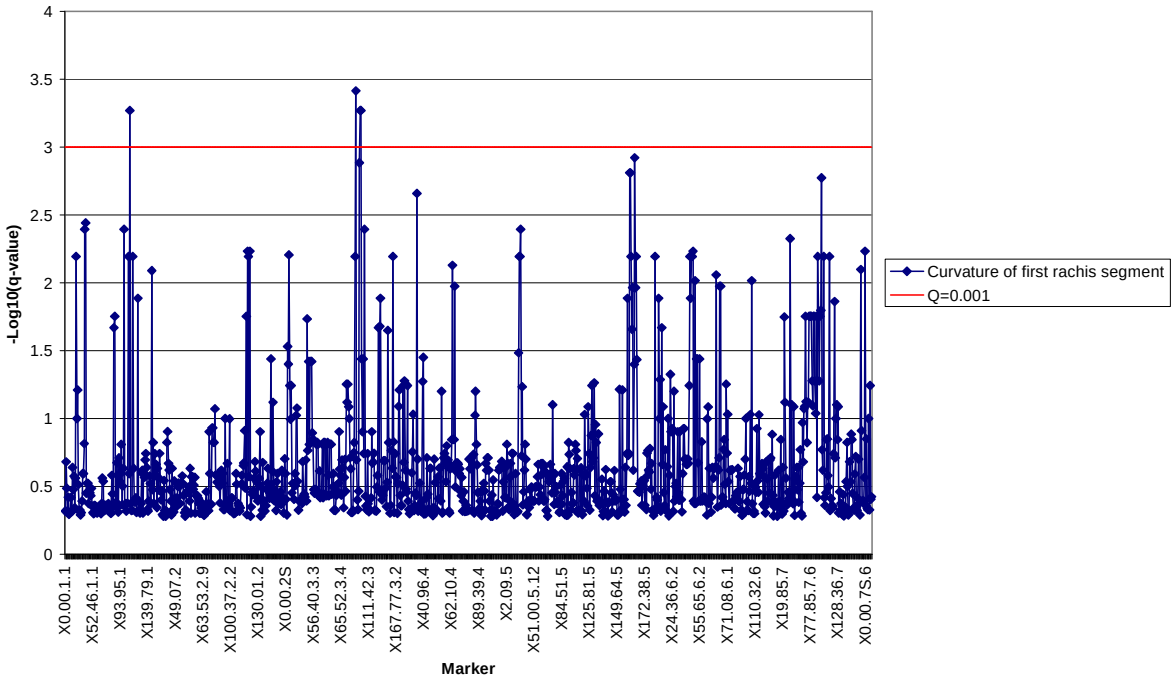


Figure 12-23. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

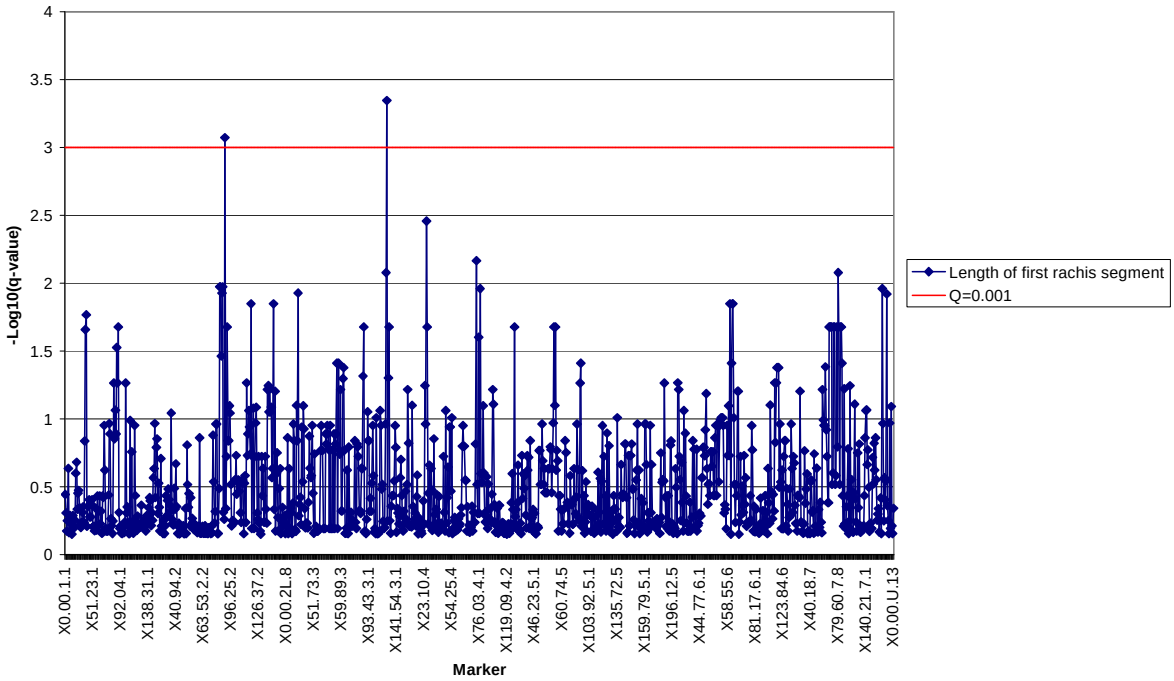


Figure 12-24. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

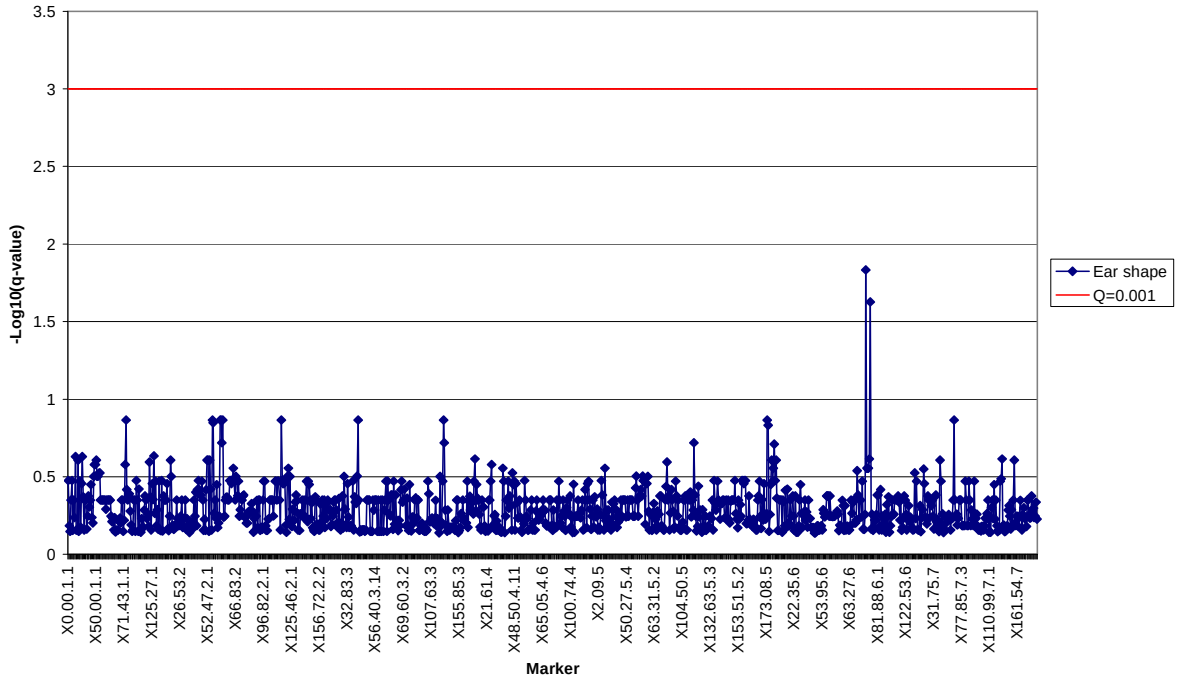


Figure 12-25. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

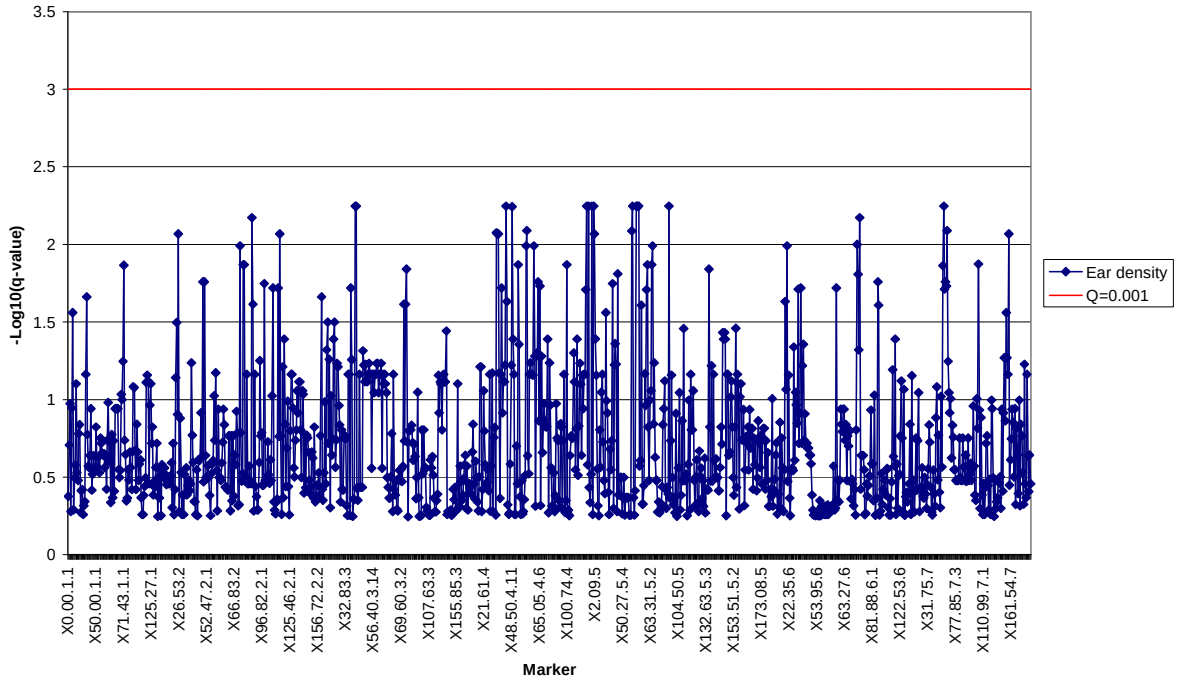


Figure 12-26. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

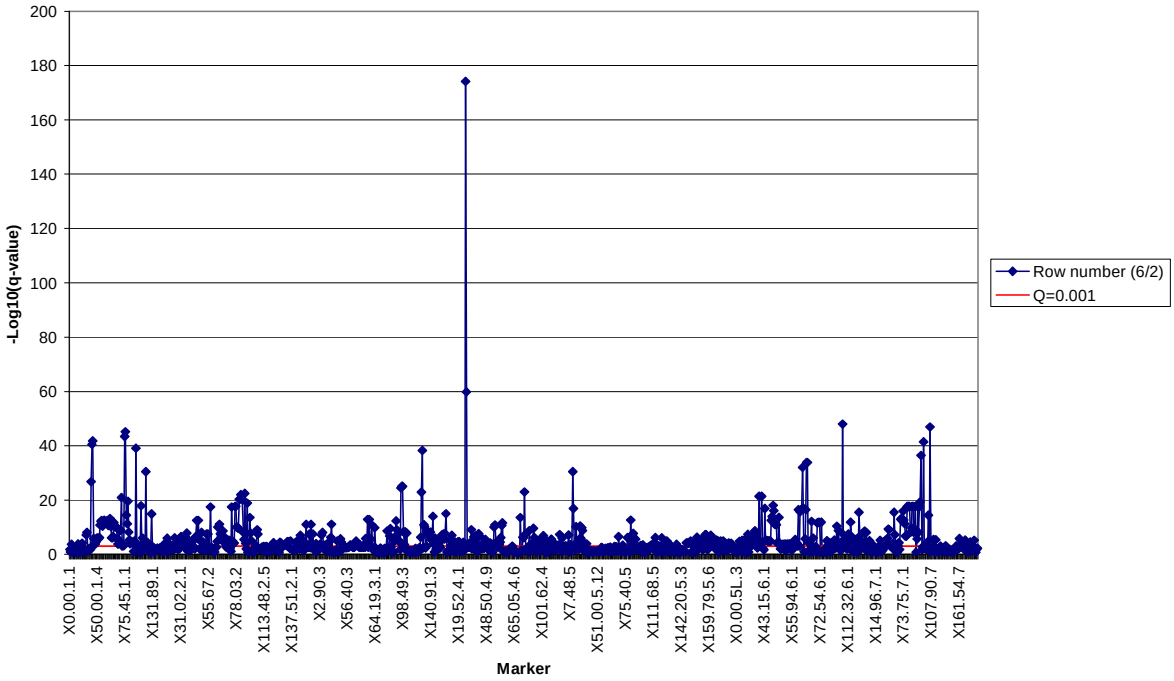


Figure 12-27. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

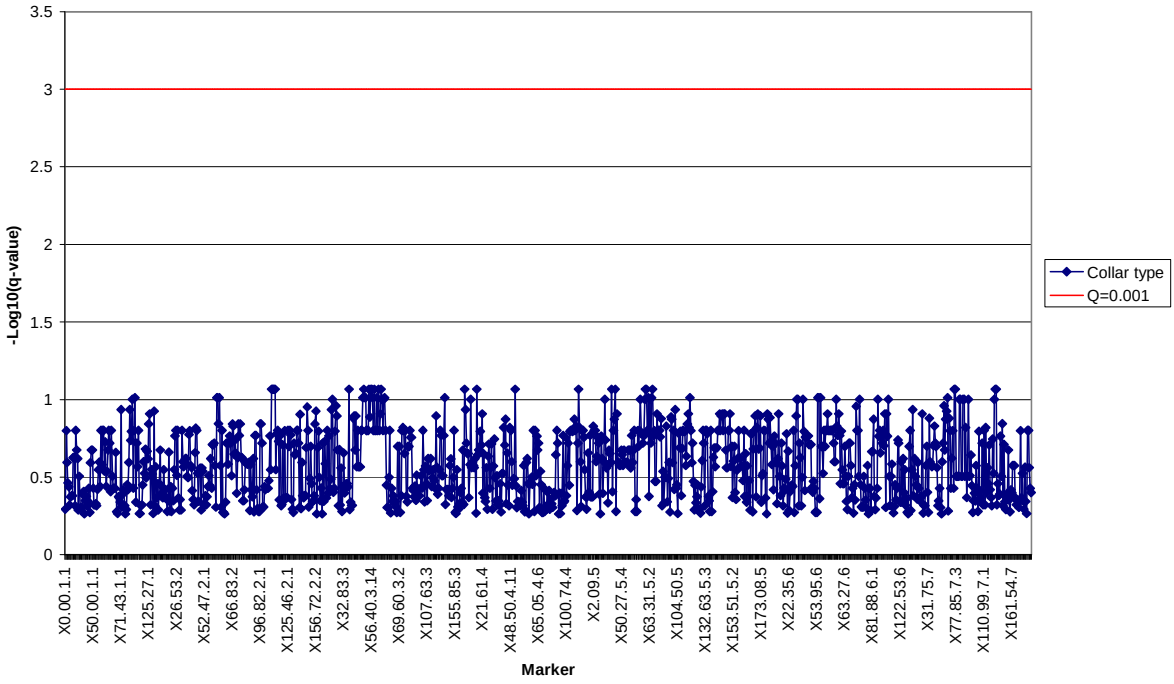


Figure 12-28. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

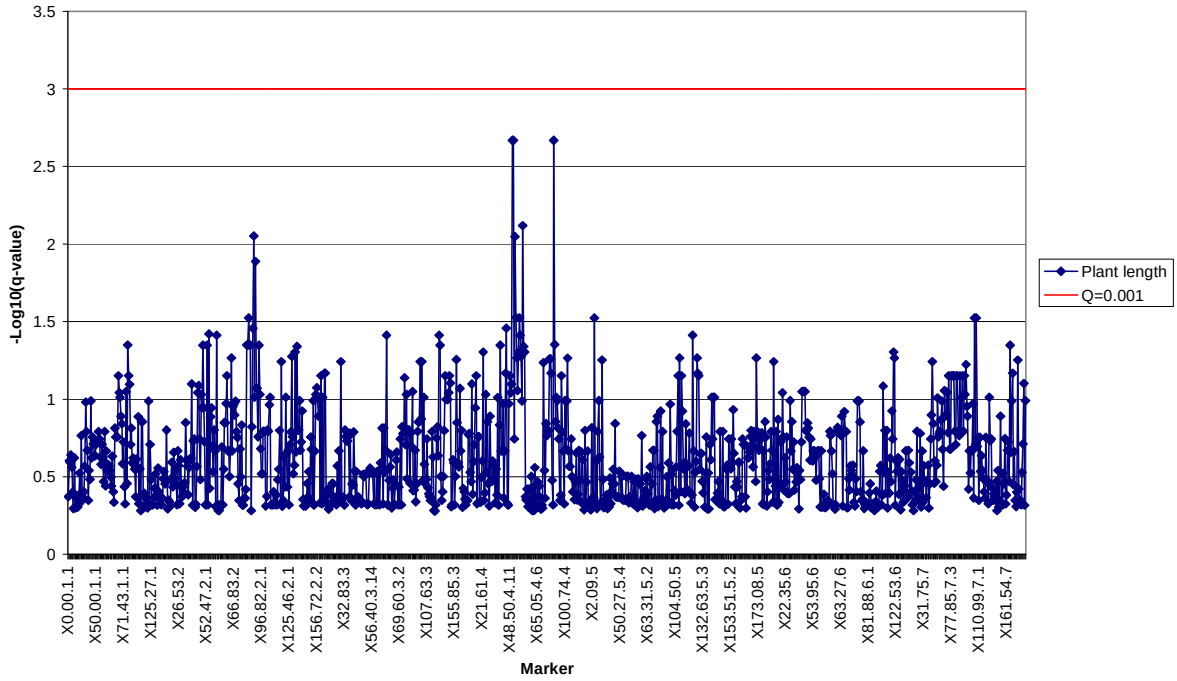


Figure 12-29. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

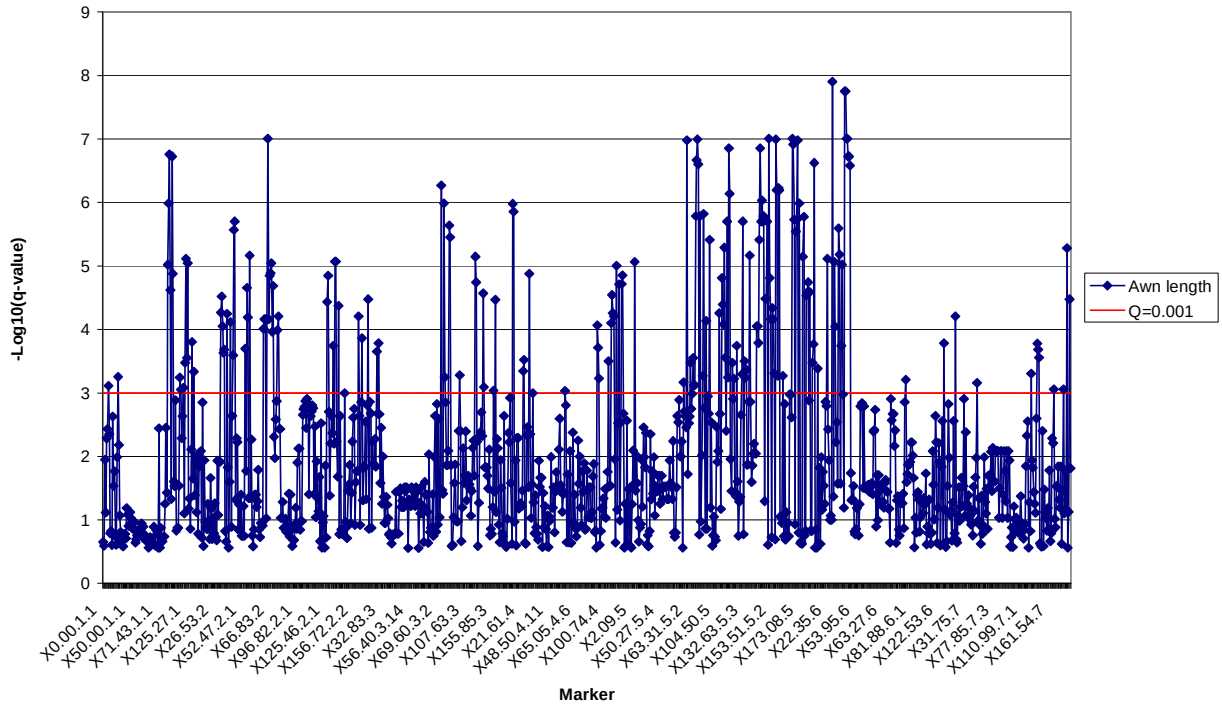


Figure 12-30. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

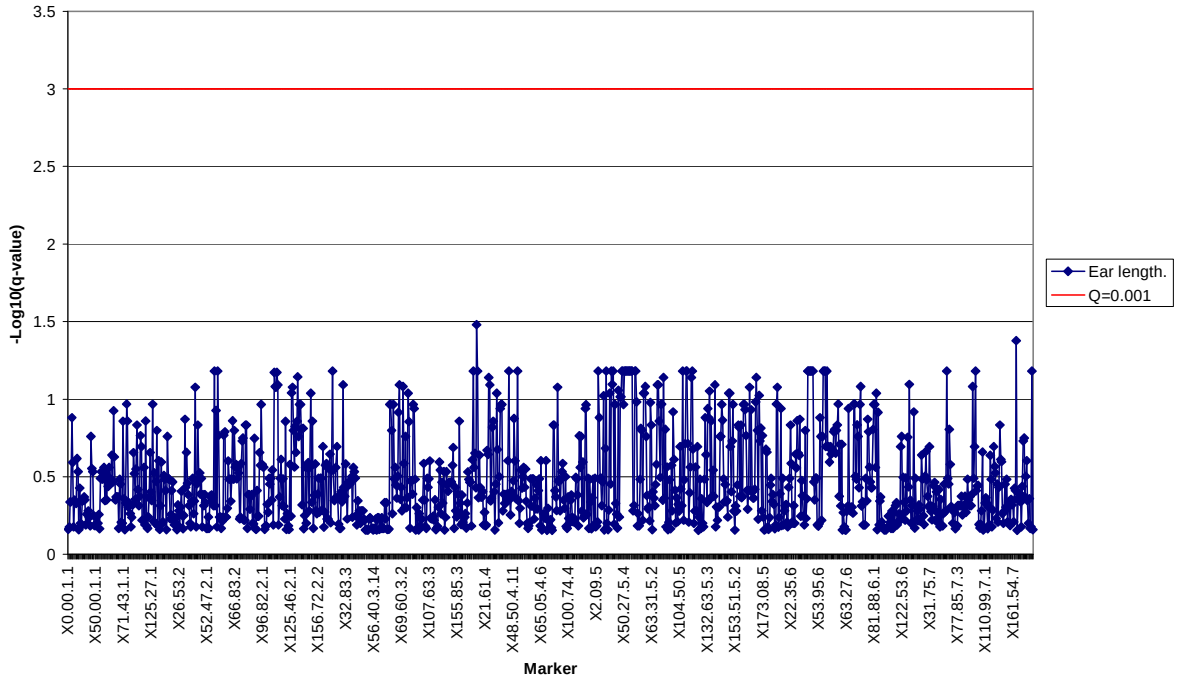


Figure 12-31. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

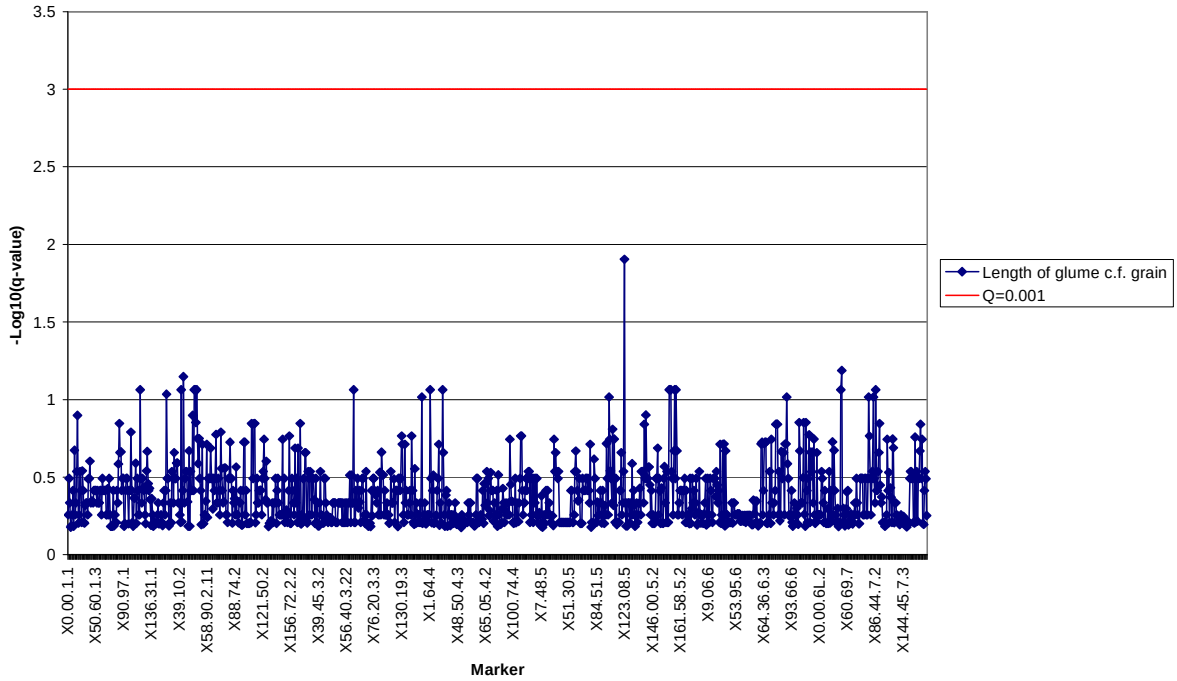


Figure 12-32. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

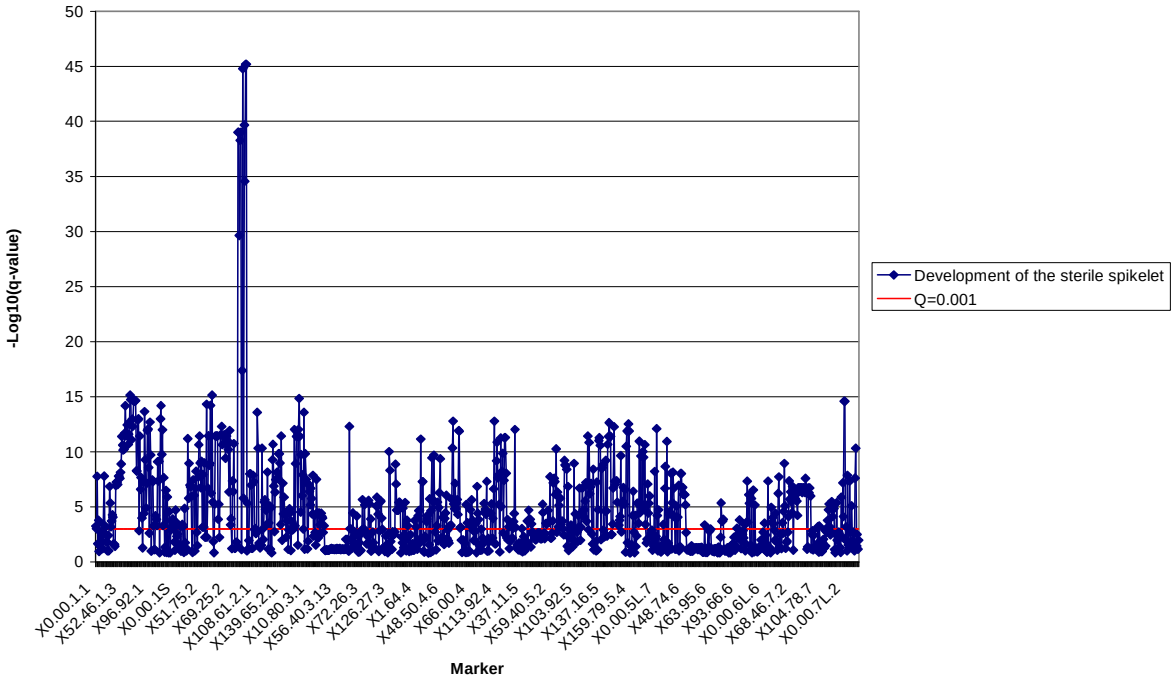


Figure 12-33. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

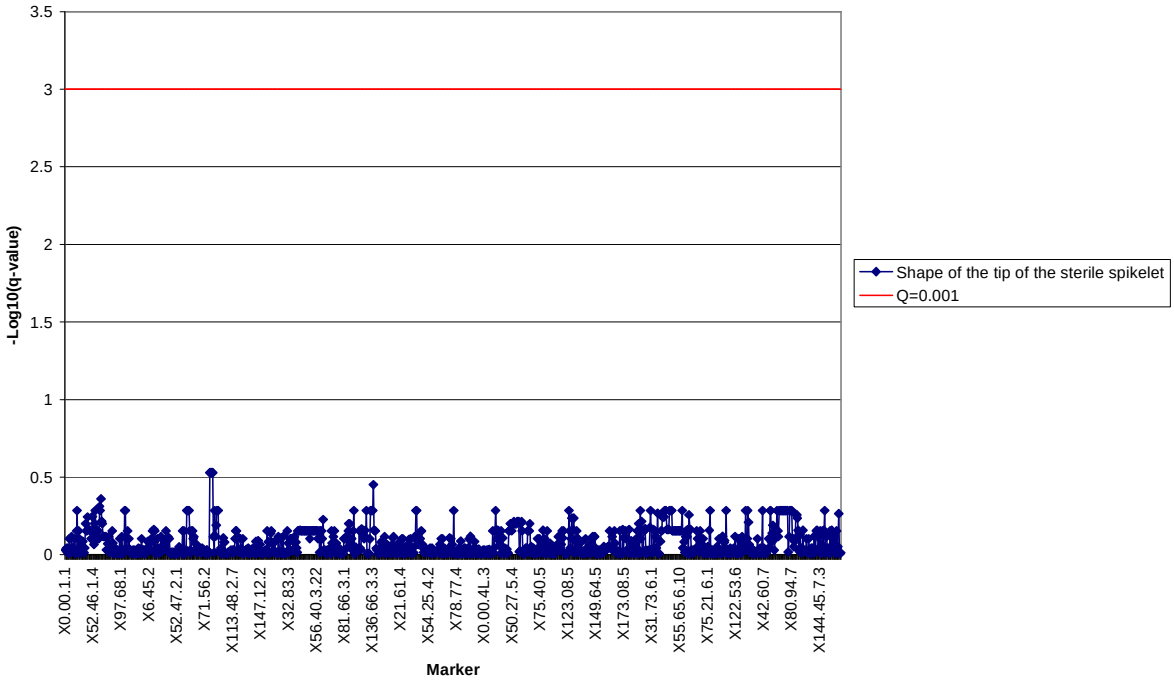


Figure 12-34. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

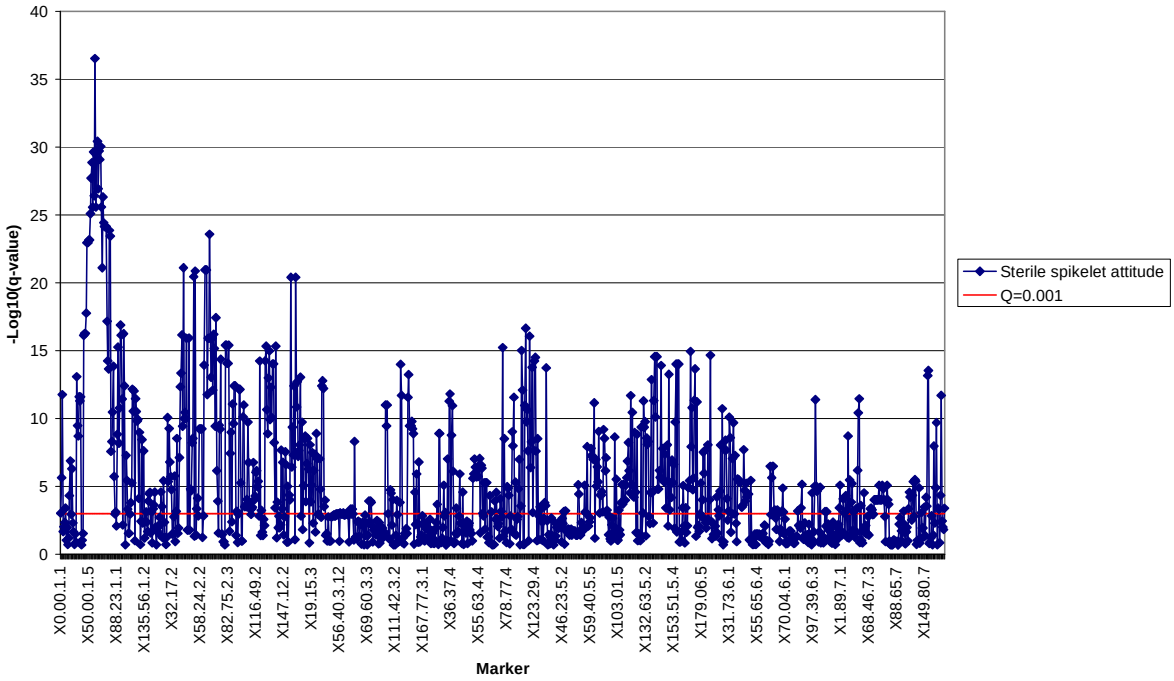


Figure 12-35. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

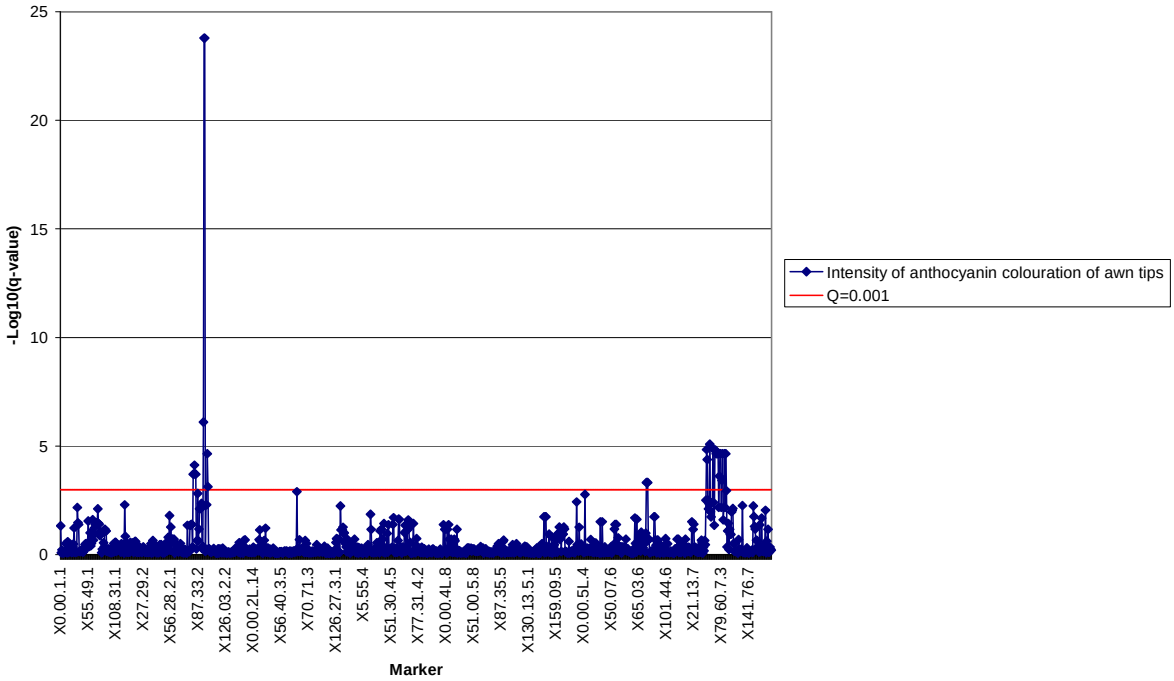


Figure 12-36. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

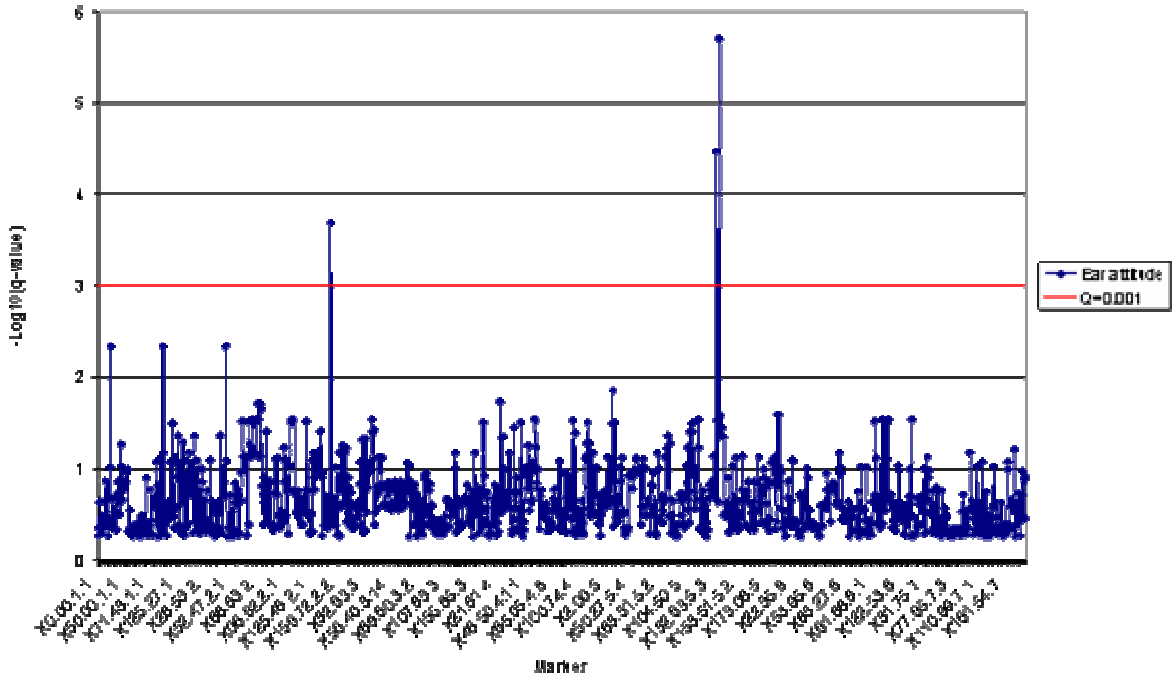


Figure 12-37. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

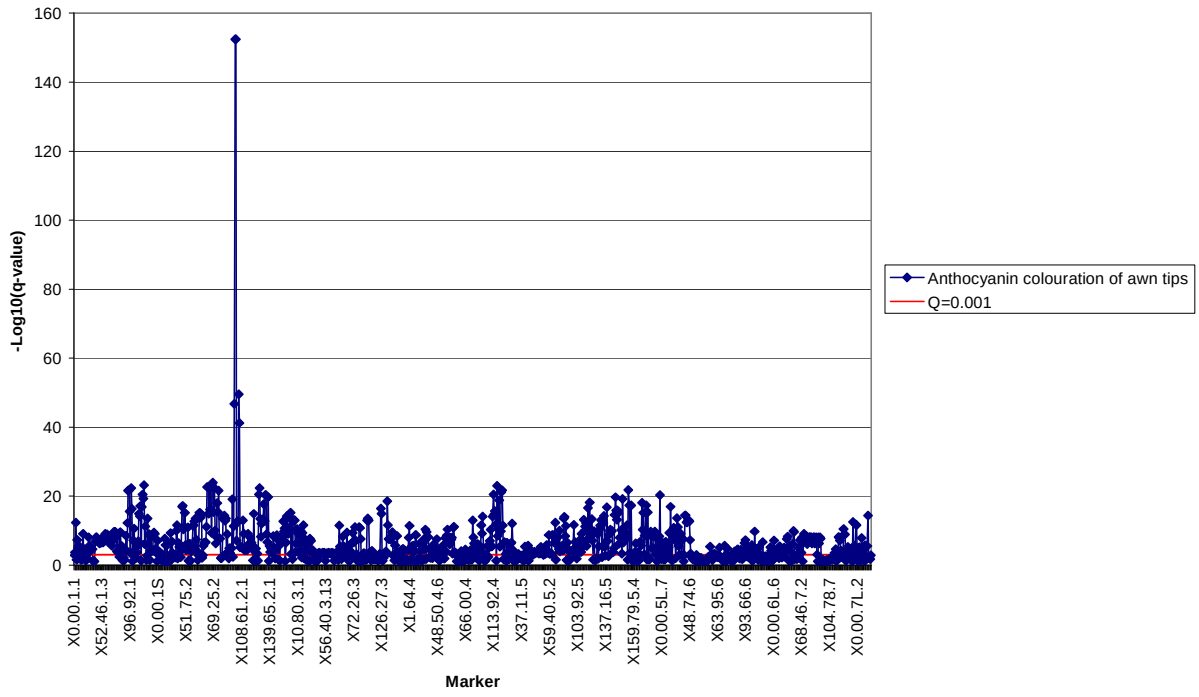


Figure 12-38. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

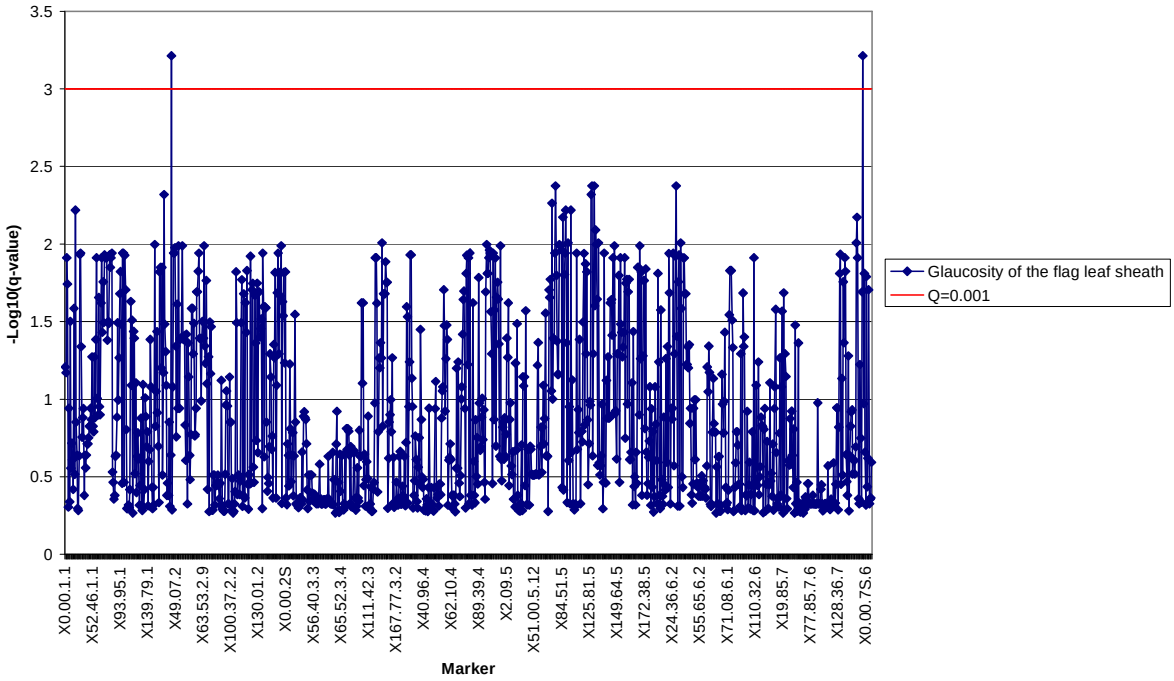


Figure 12-39. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

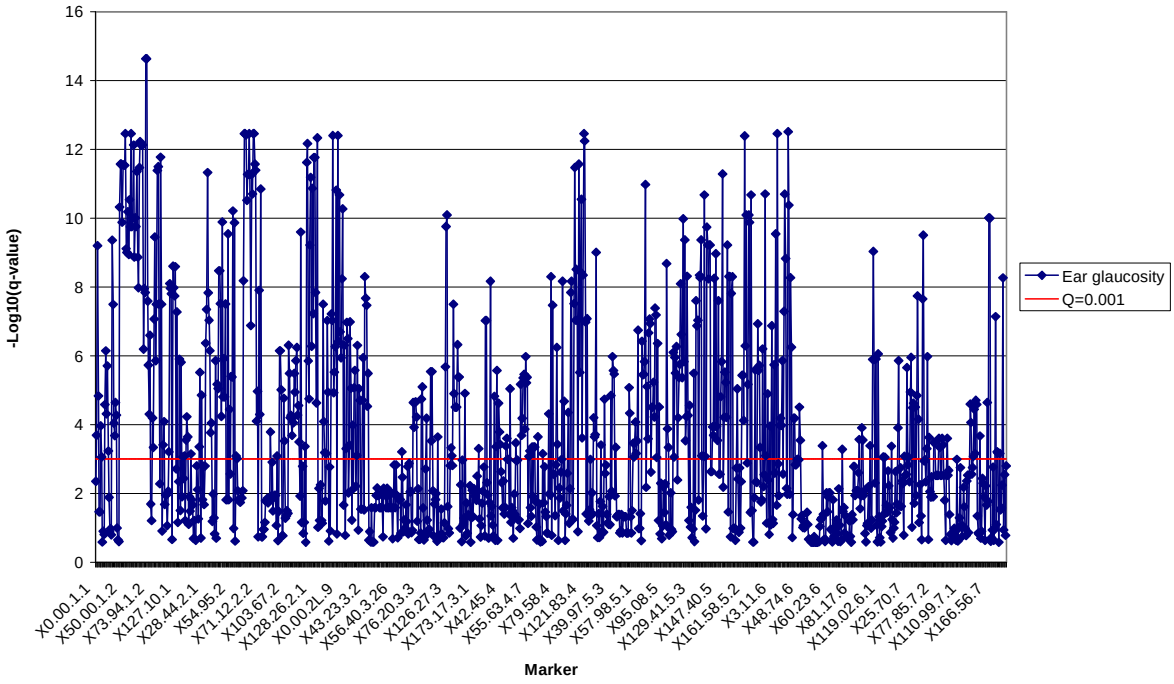


Figure 12-40. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

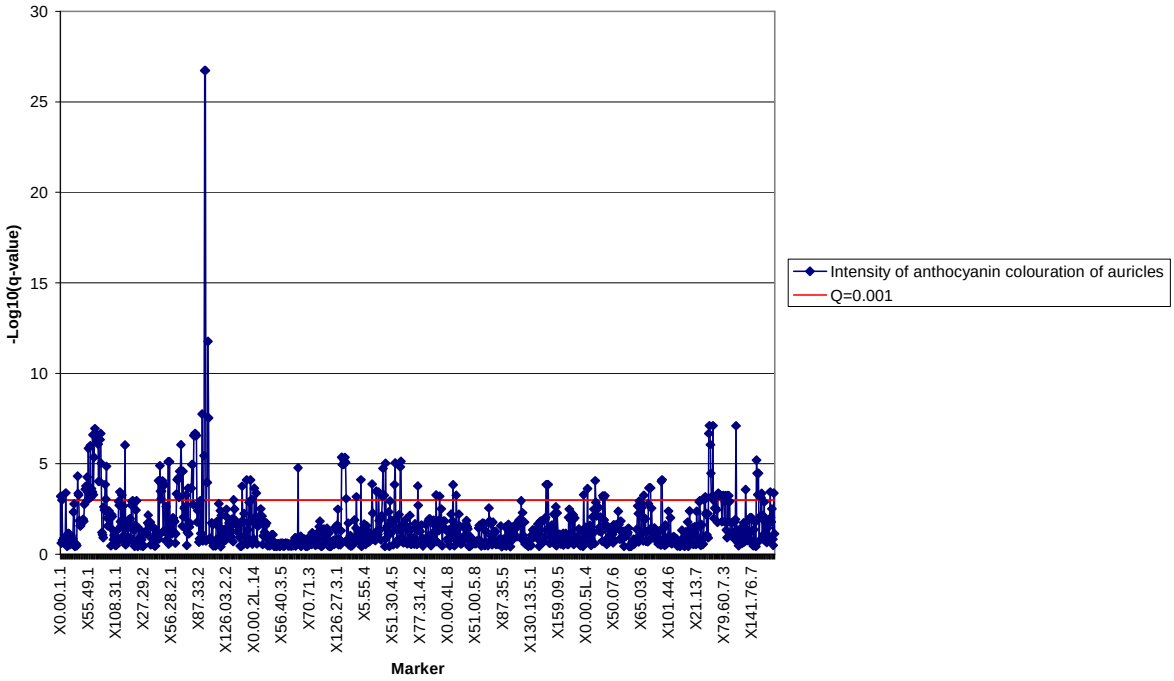


Figure 12-41. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

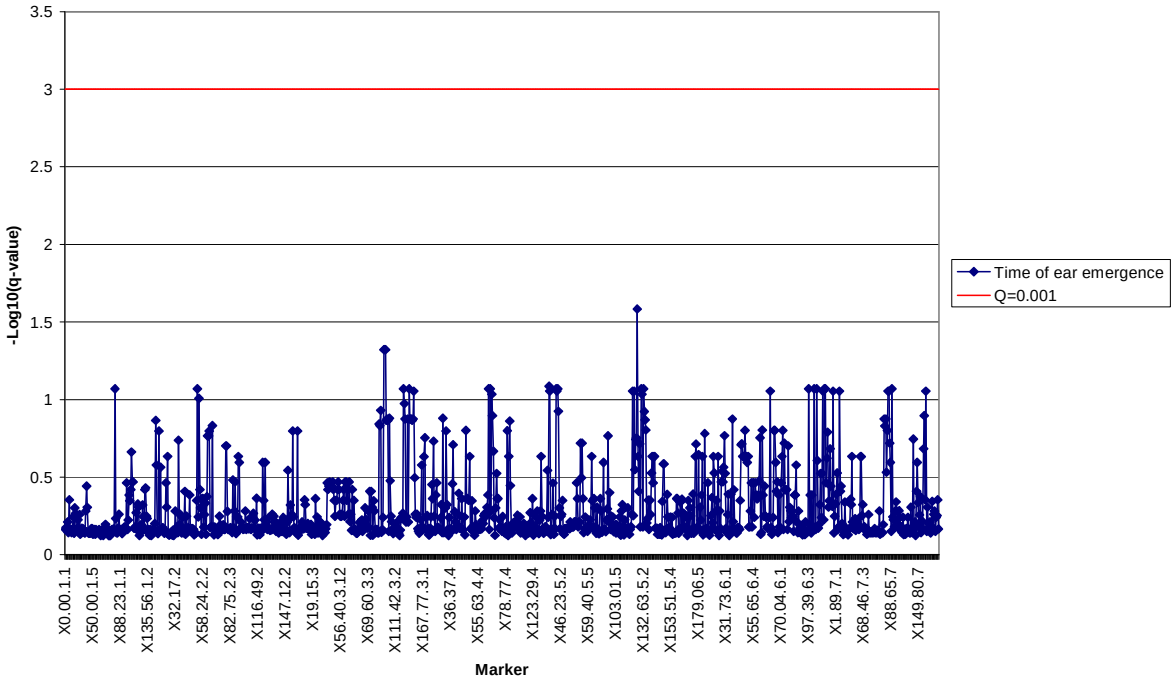


Figure 12-42. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

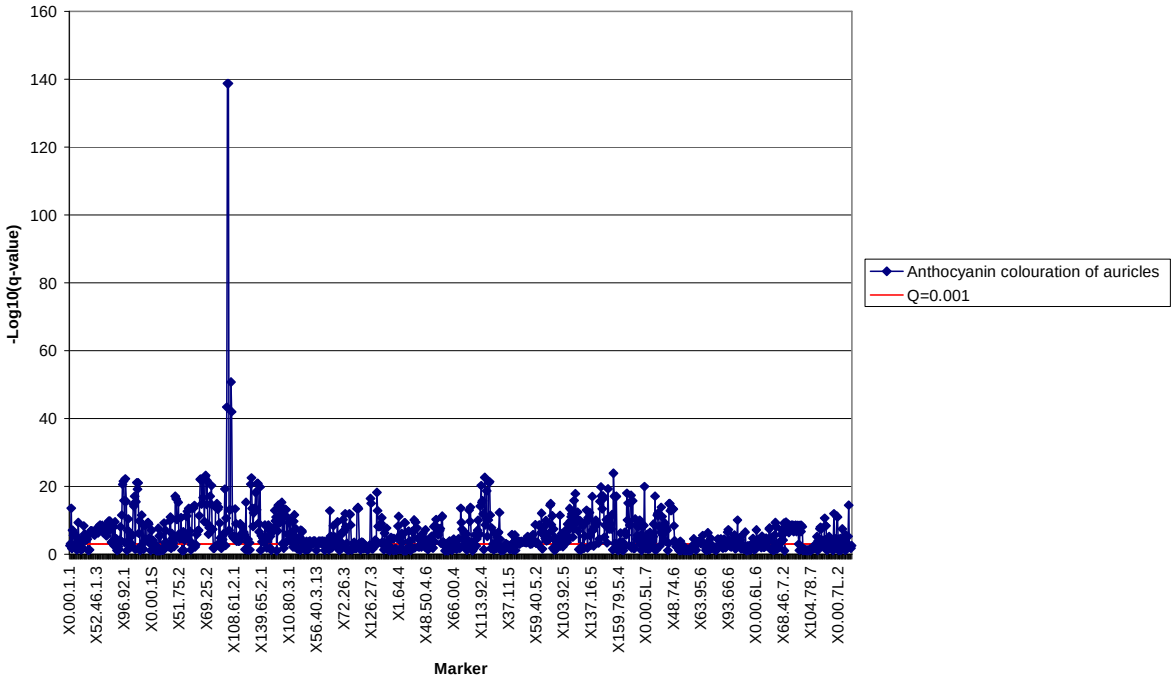


Figure 12-43. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

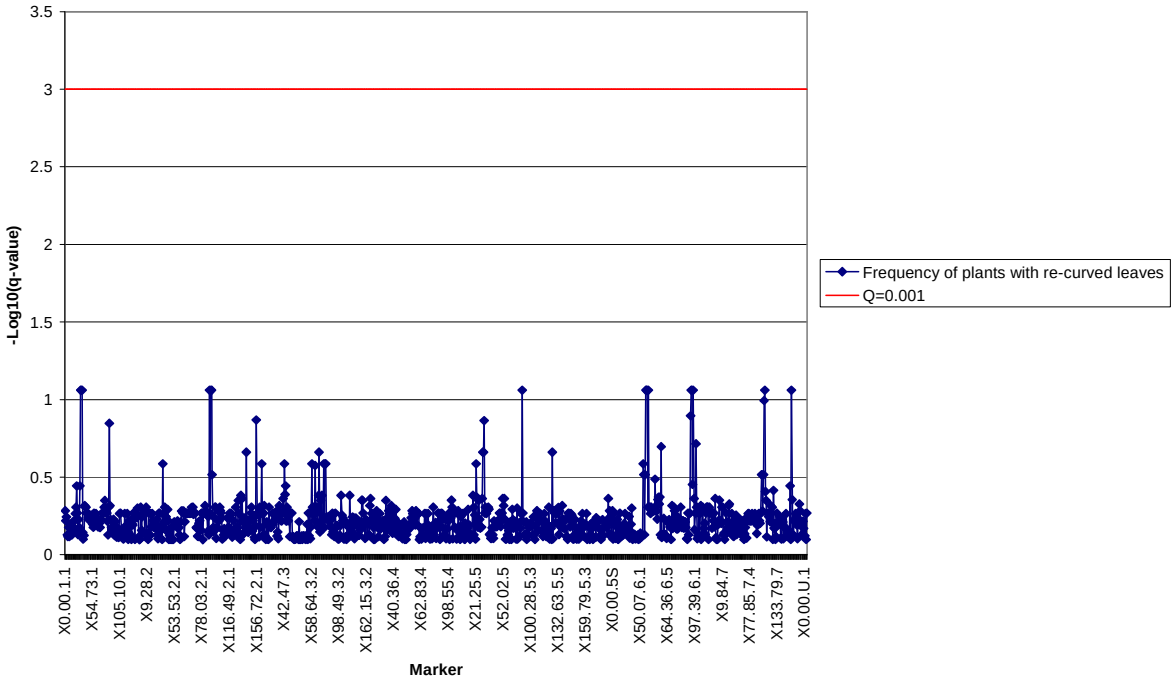


Figure 12-44. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

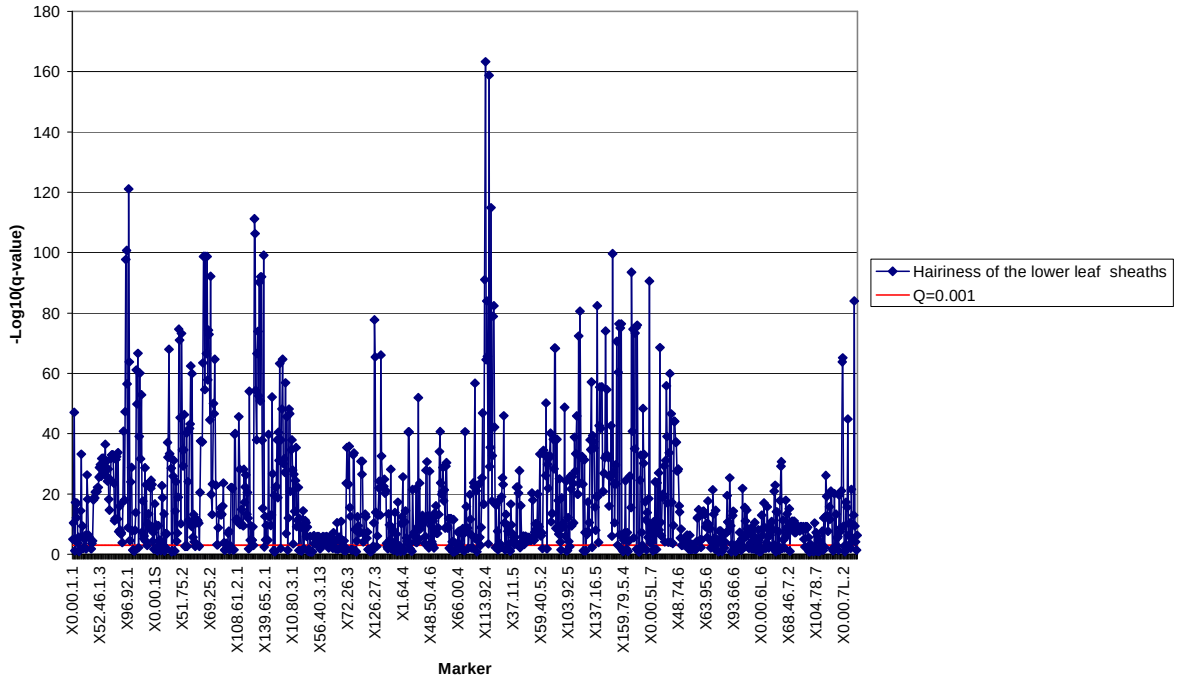


Figure 12-45. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

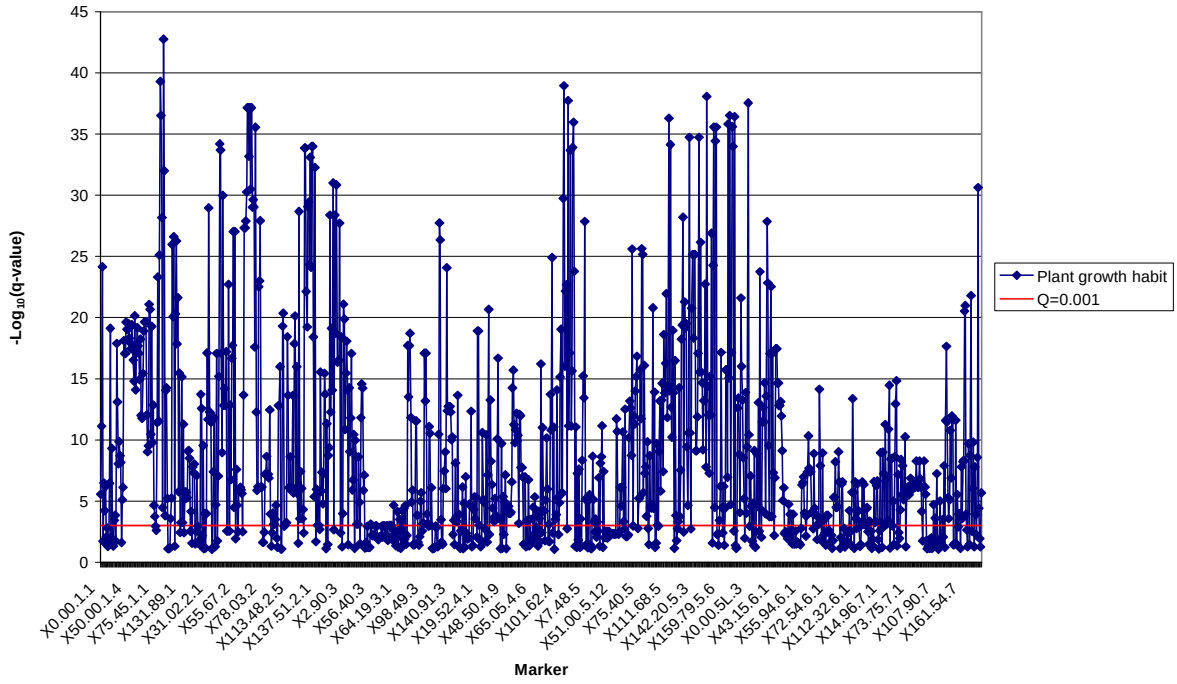


Figure 12-46. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

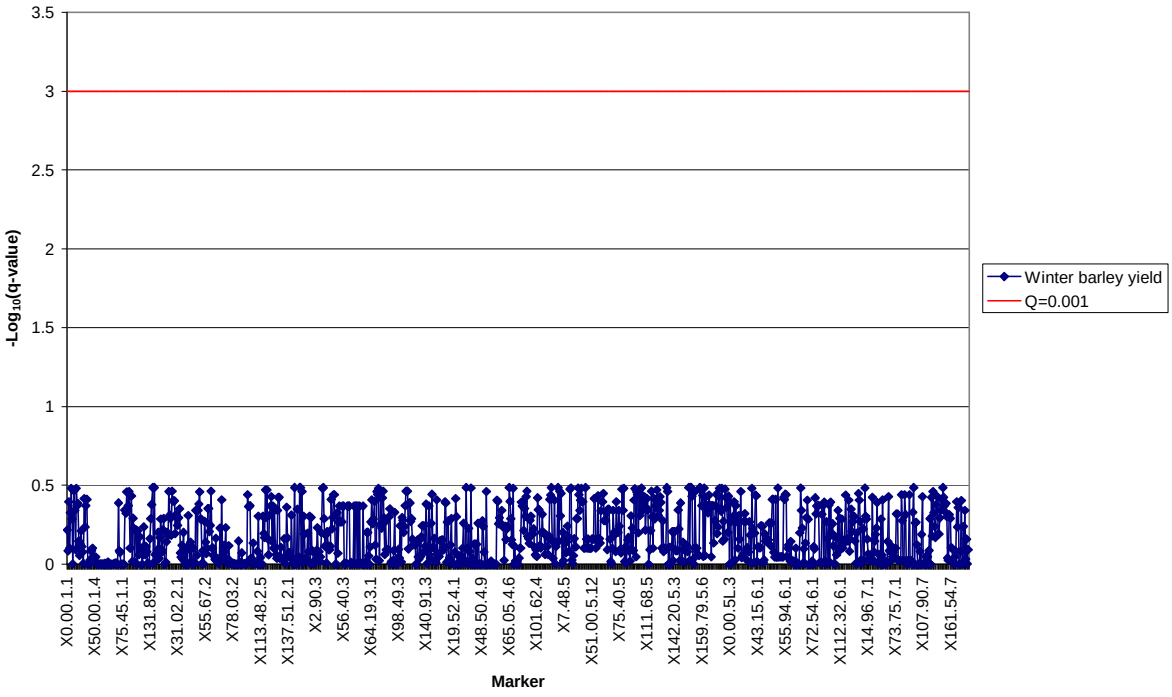


Figure 12-47. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

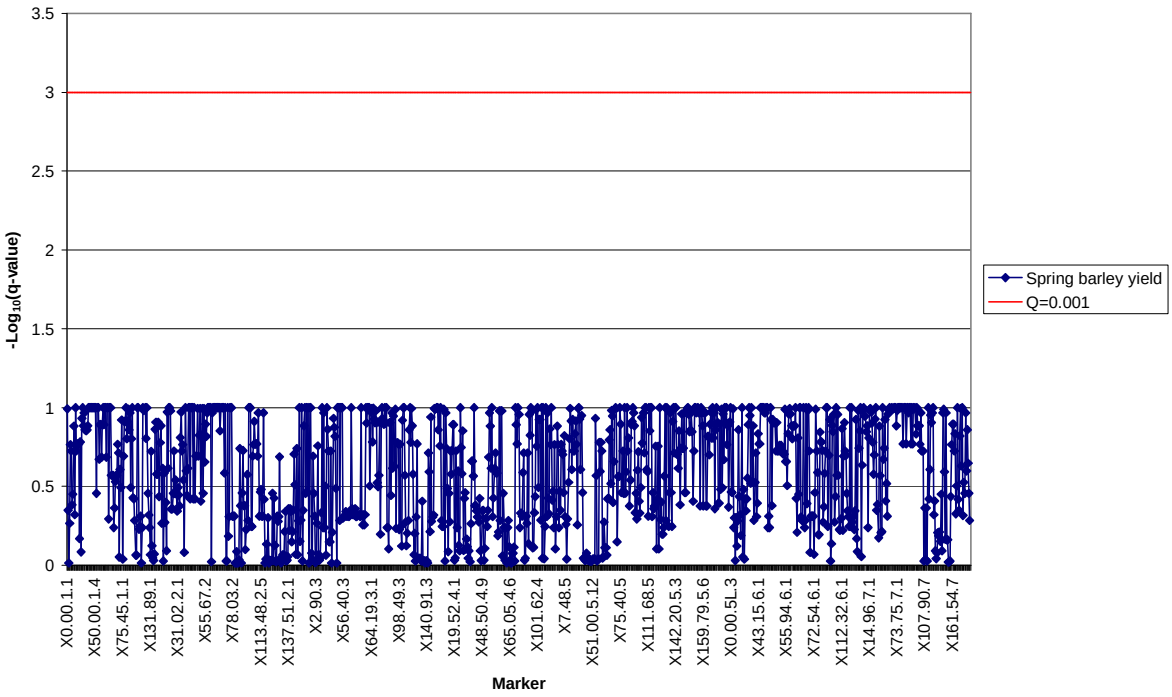


Figure 12-48. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

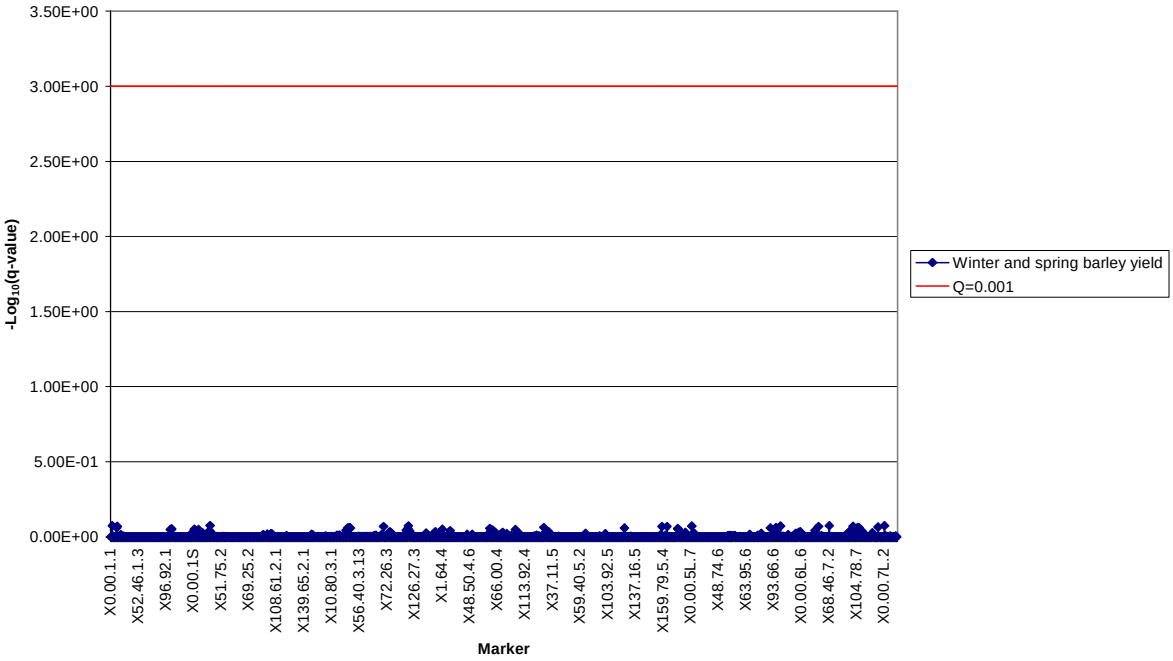


Figure 12-49. Association by uncorrected regression. Markers arranged in genetic map order. Trait as indicated in legend.

12.8 Appendix 8. Association plots for barley phenotypes. Association estimated following adjustment for structure using SA + GC.

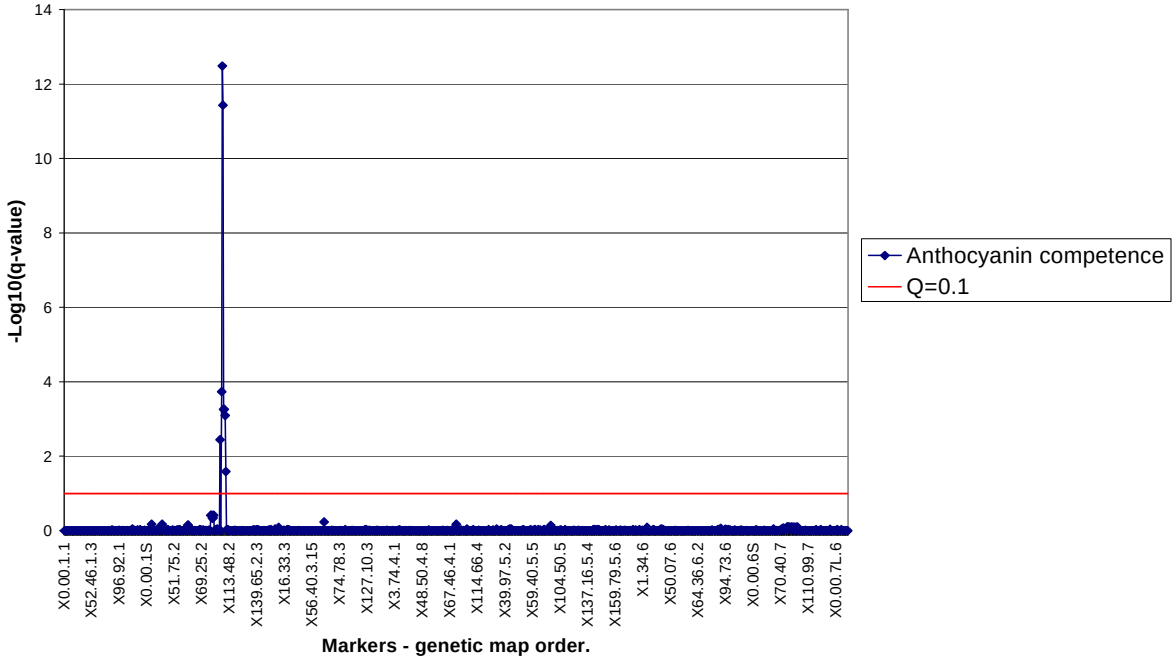


Figure 12-50. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend.

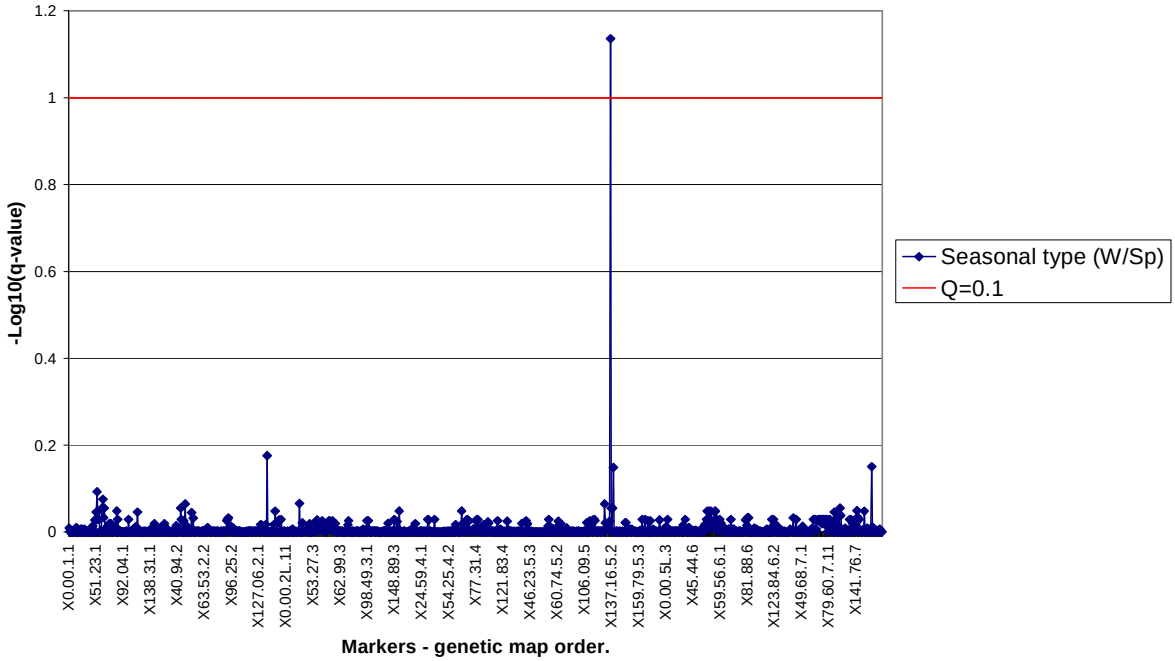


Figure 12-51. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

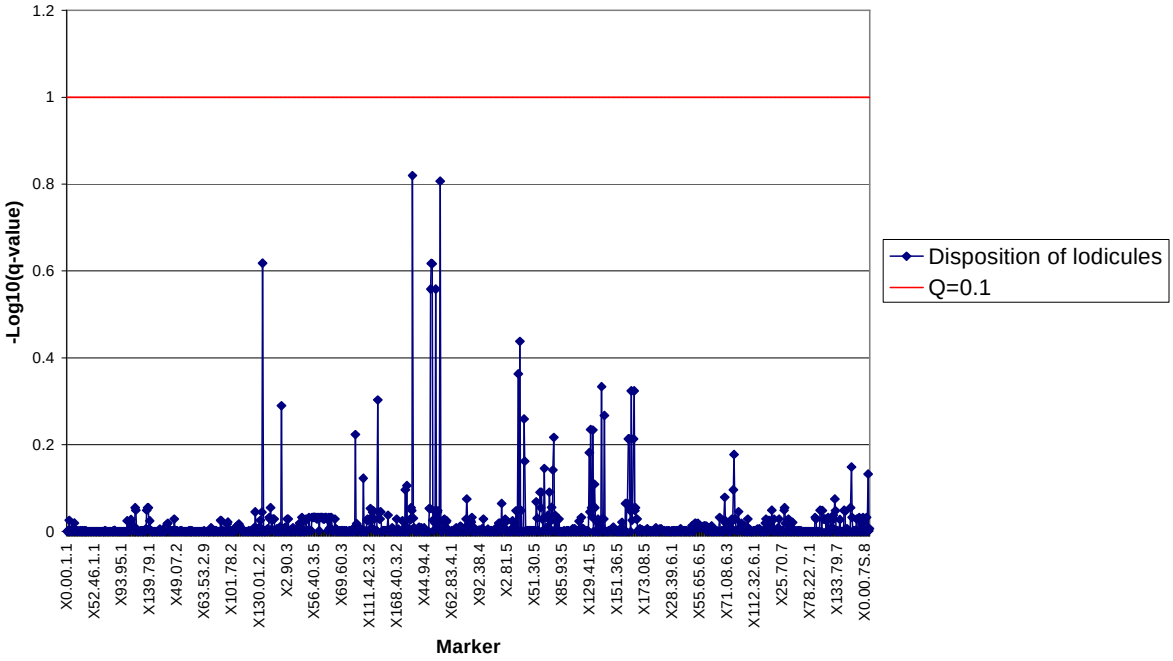


Figure 12-52. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

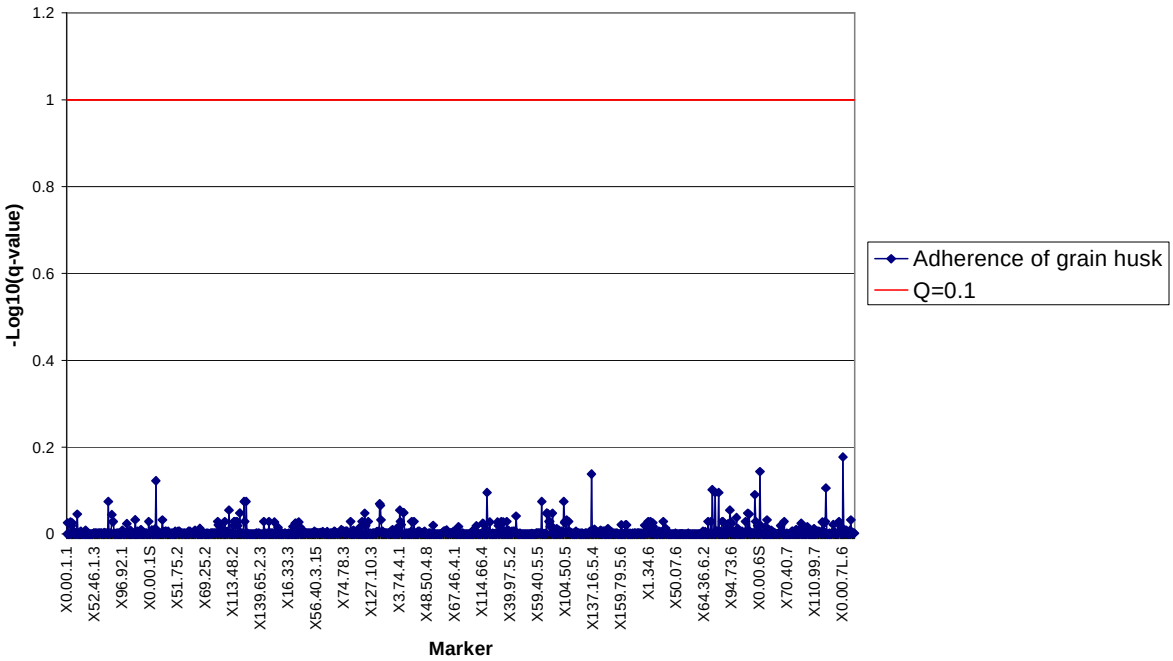


Figure 12-53. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

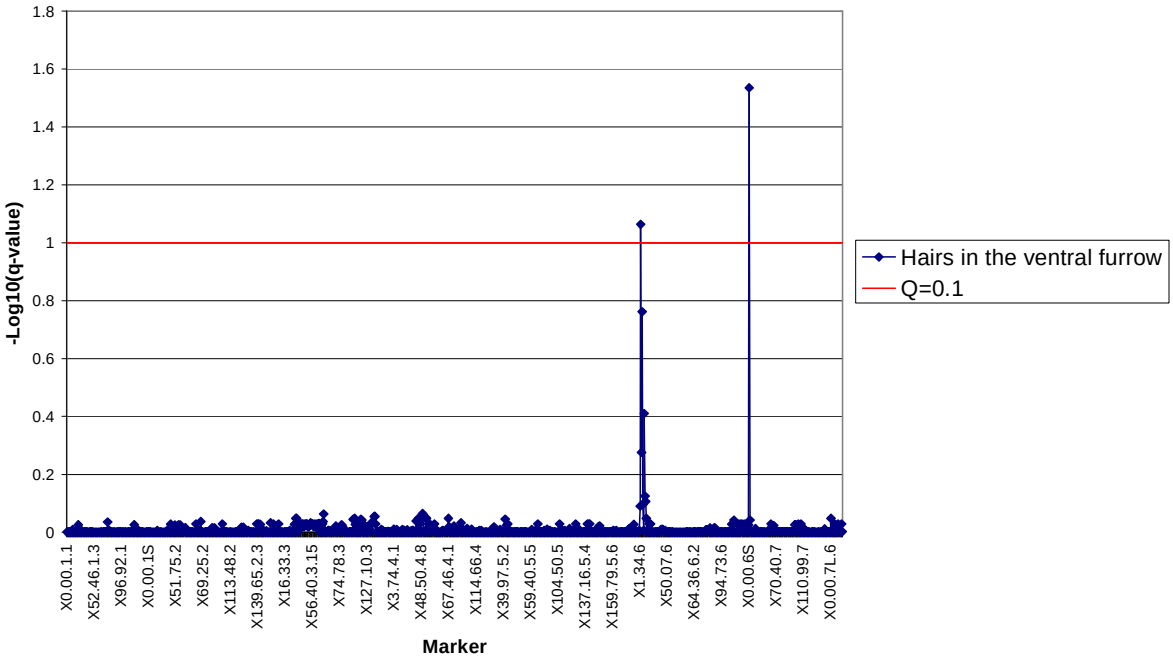


Figure 12-54. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

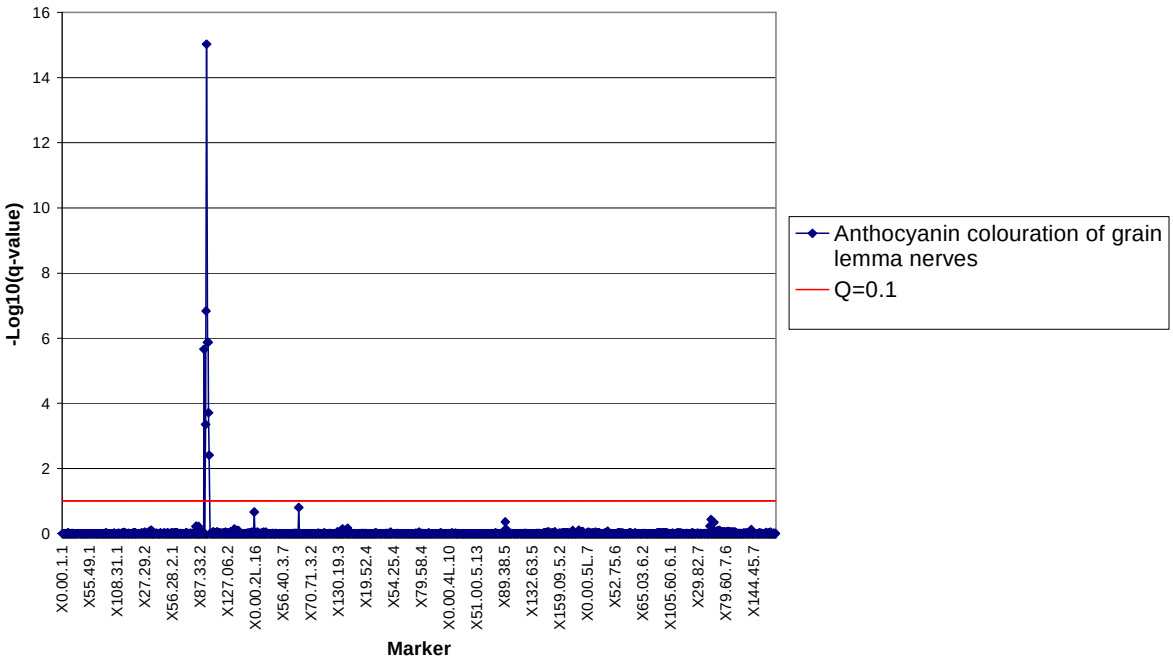


Figure 12-55. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

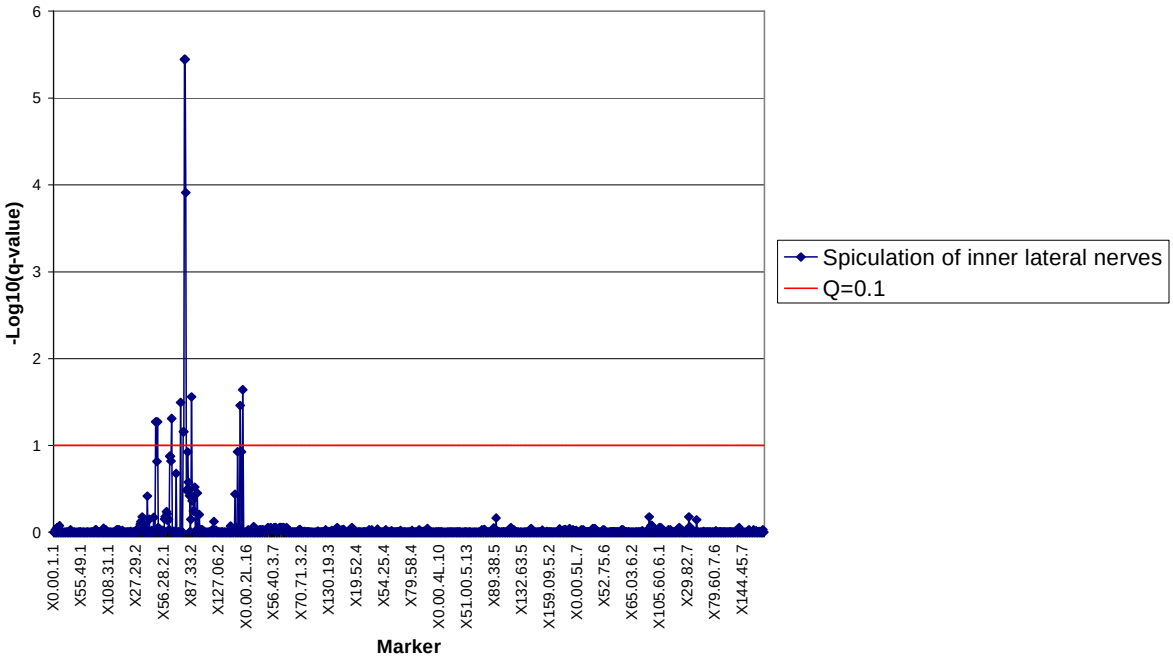


Figure 12-56. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

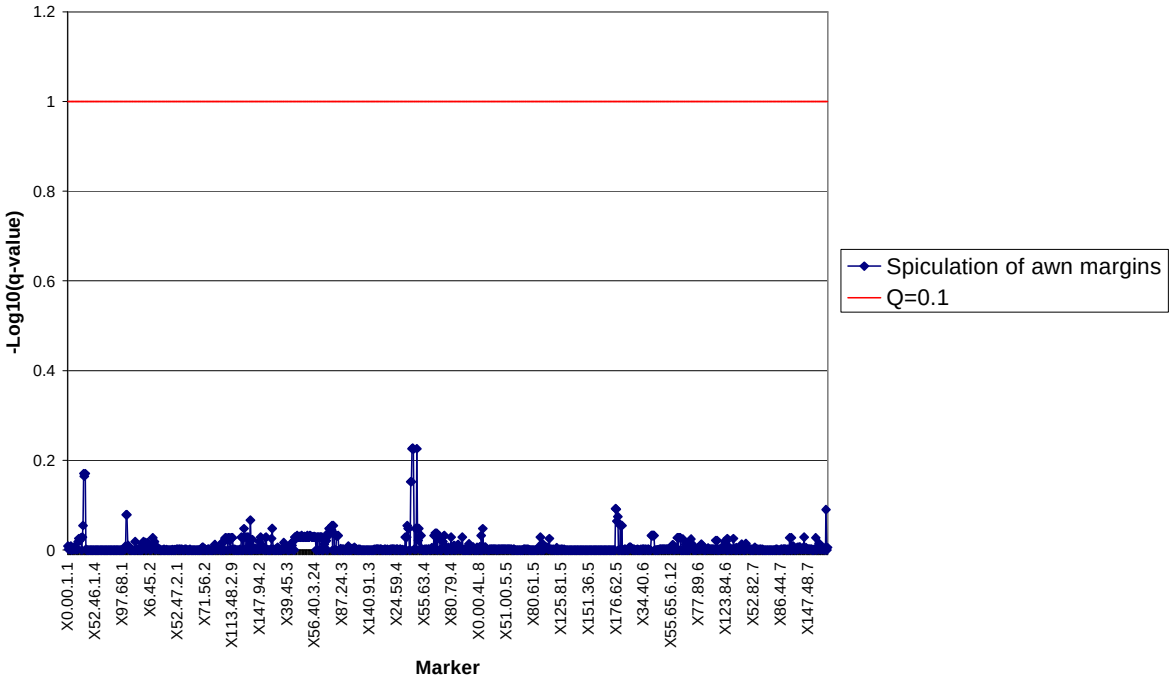


Figure 12-57. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

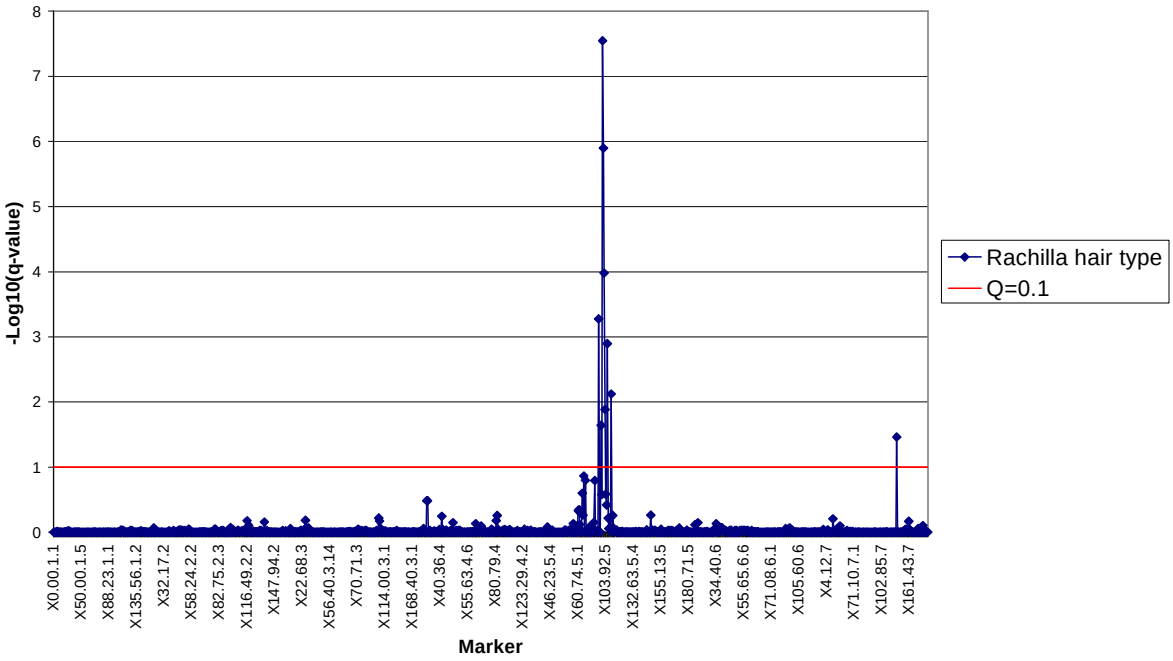


Figure 12-58. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

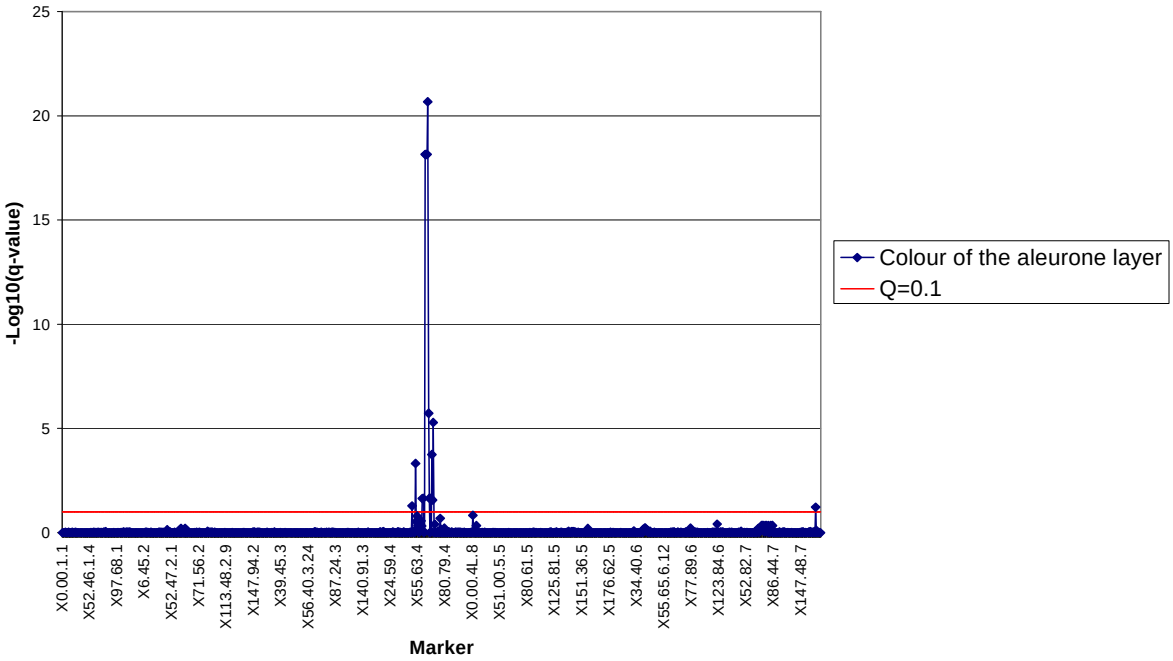


Figure 12-59. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

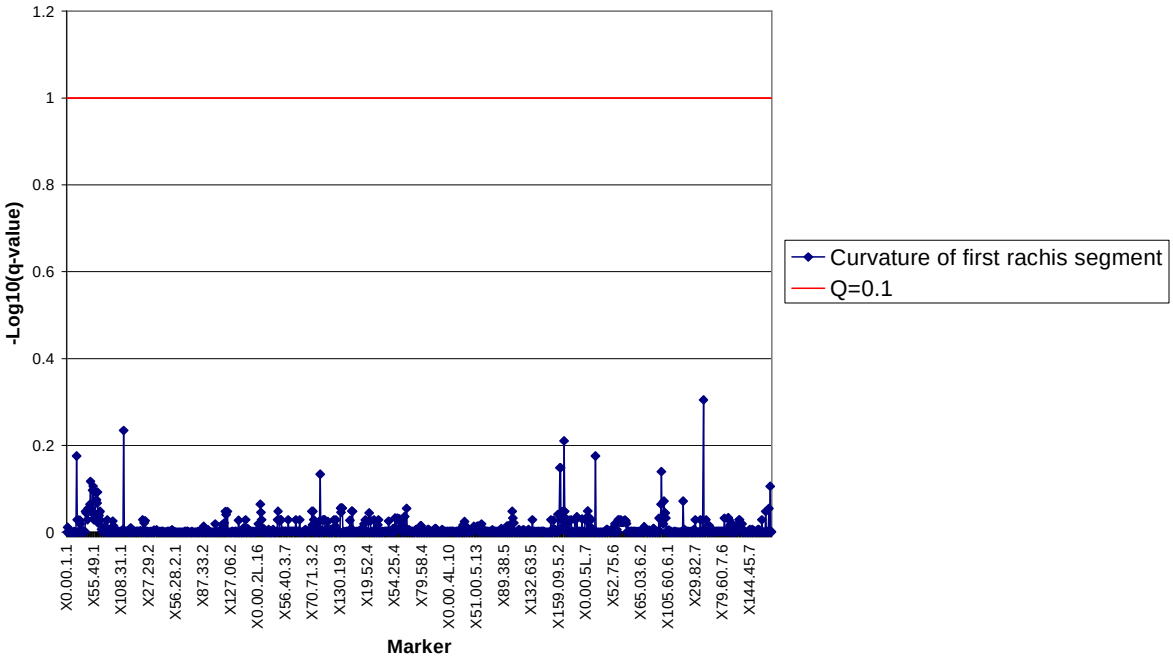


Figure 12-60. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

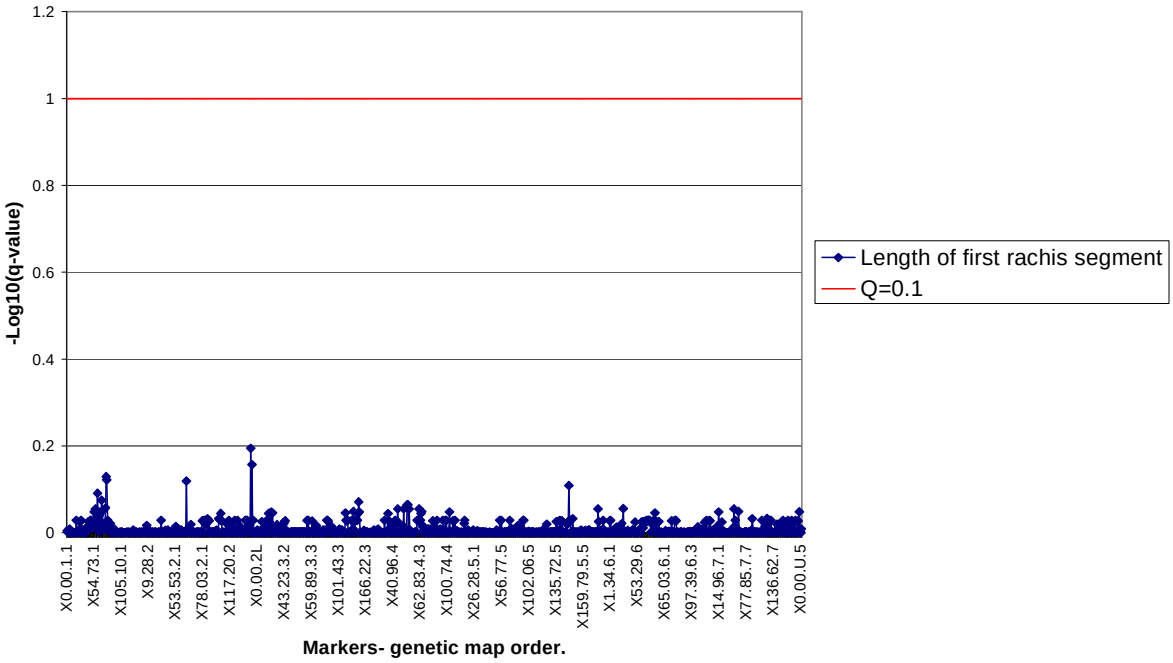


Figure 12-61. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

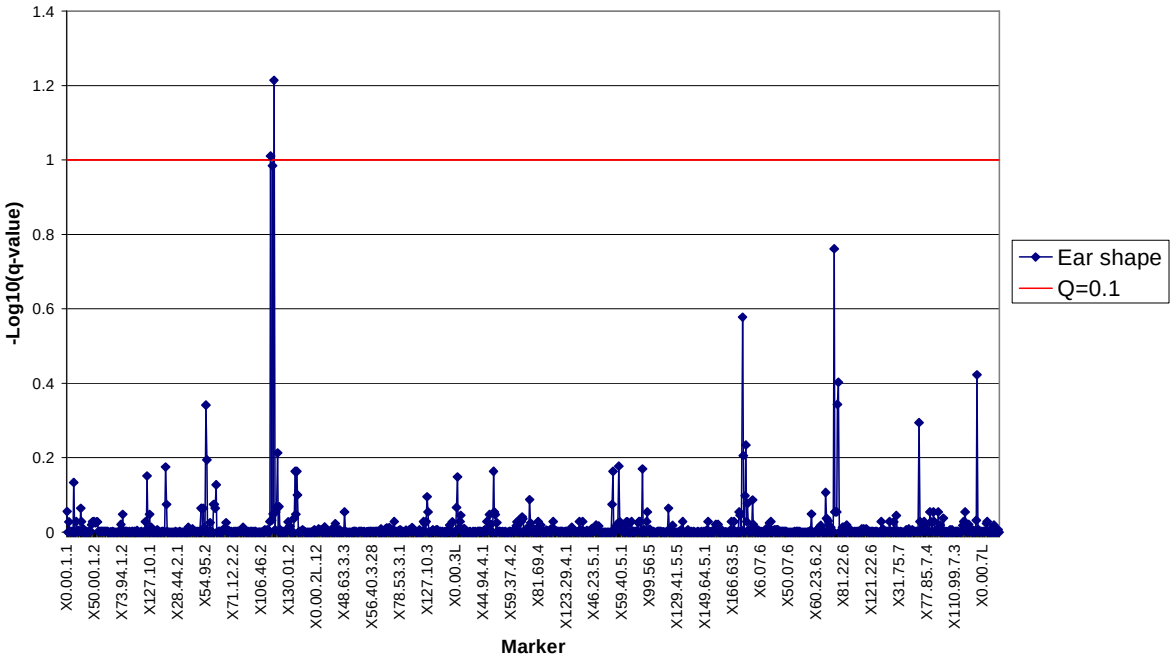


Figure 12-62. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

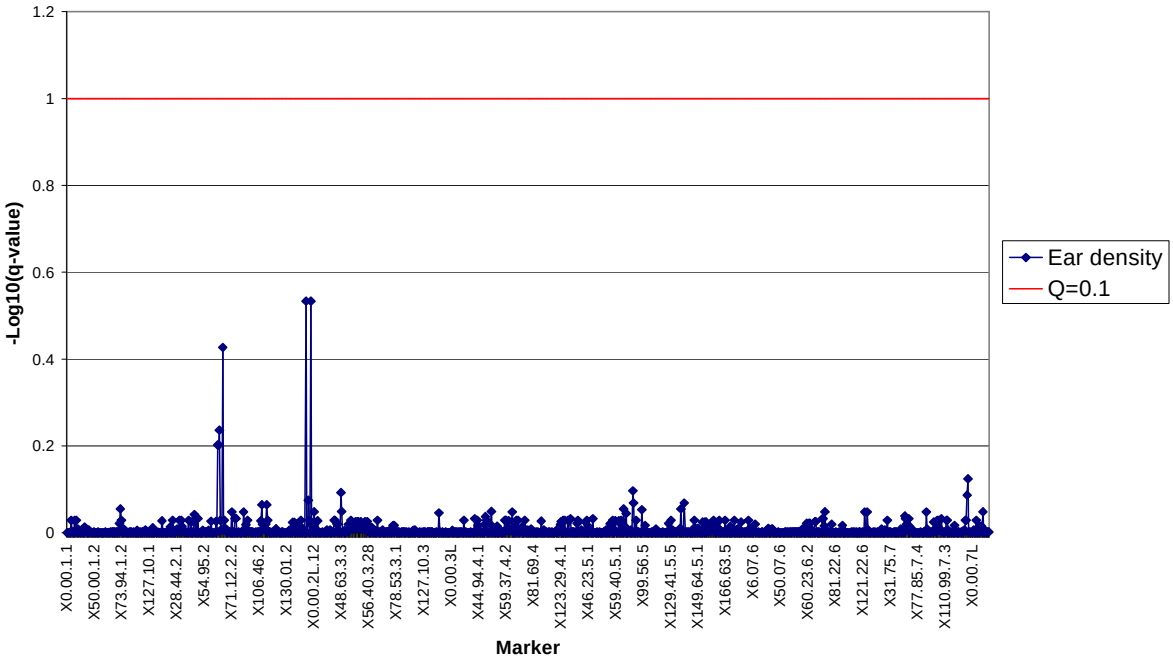


Figure 12-63. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

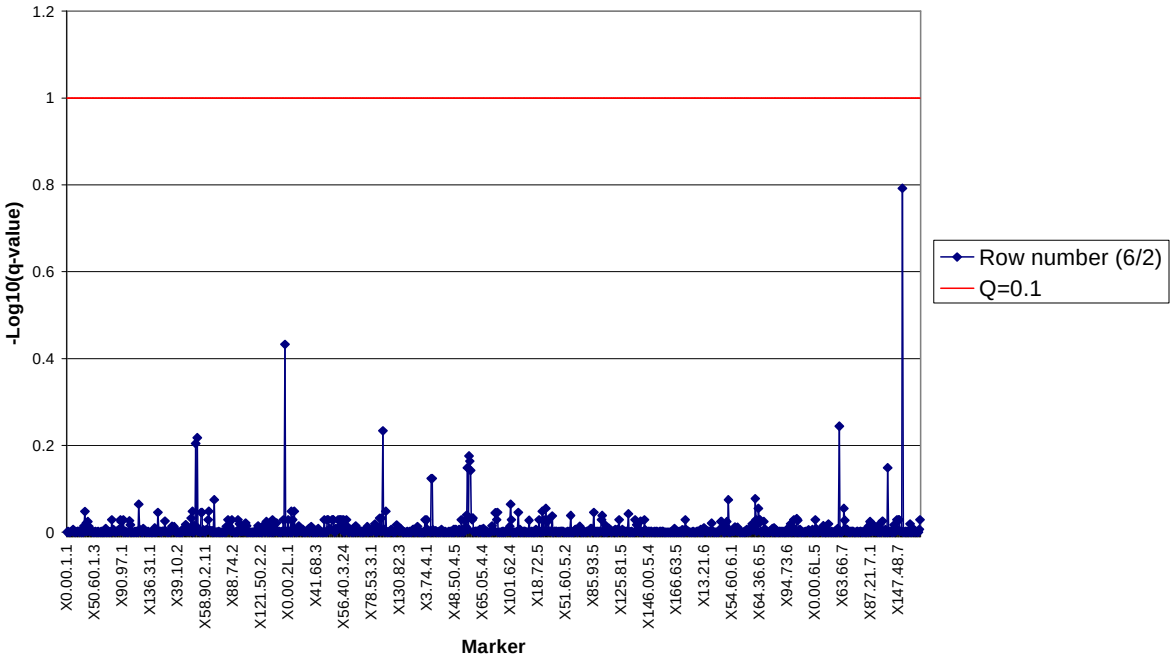


Figure 12-64. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

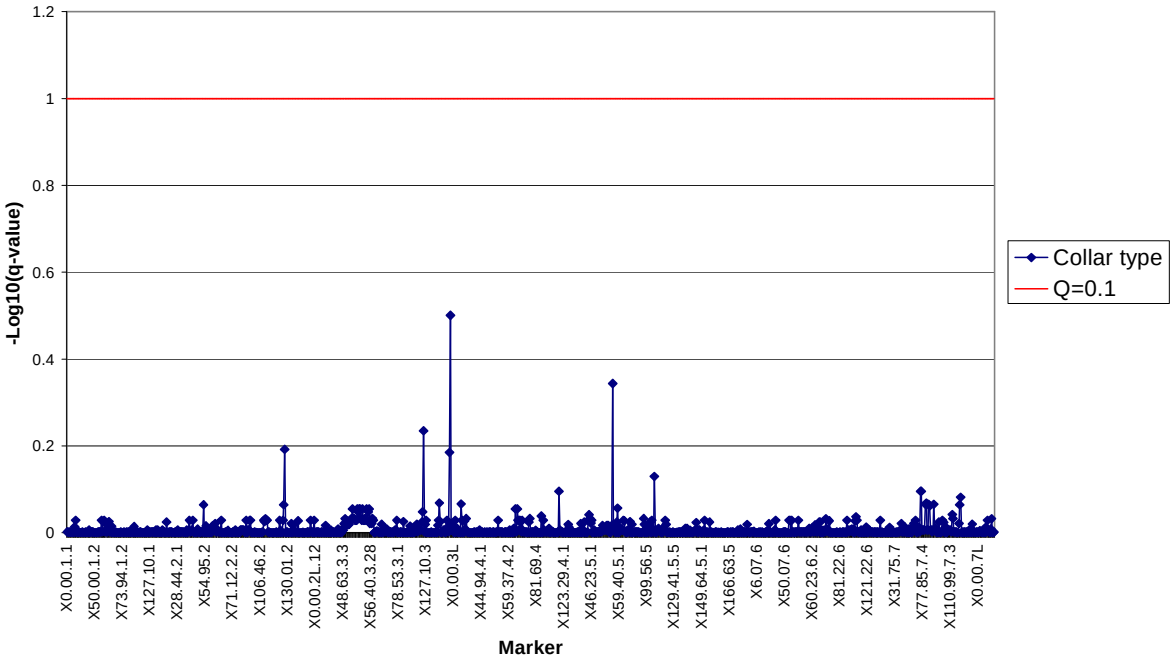


Figure 12-65. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

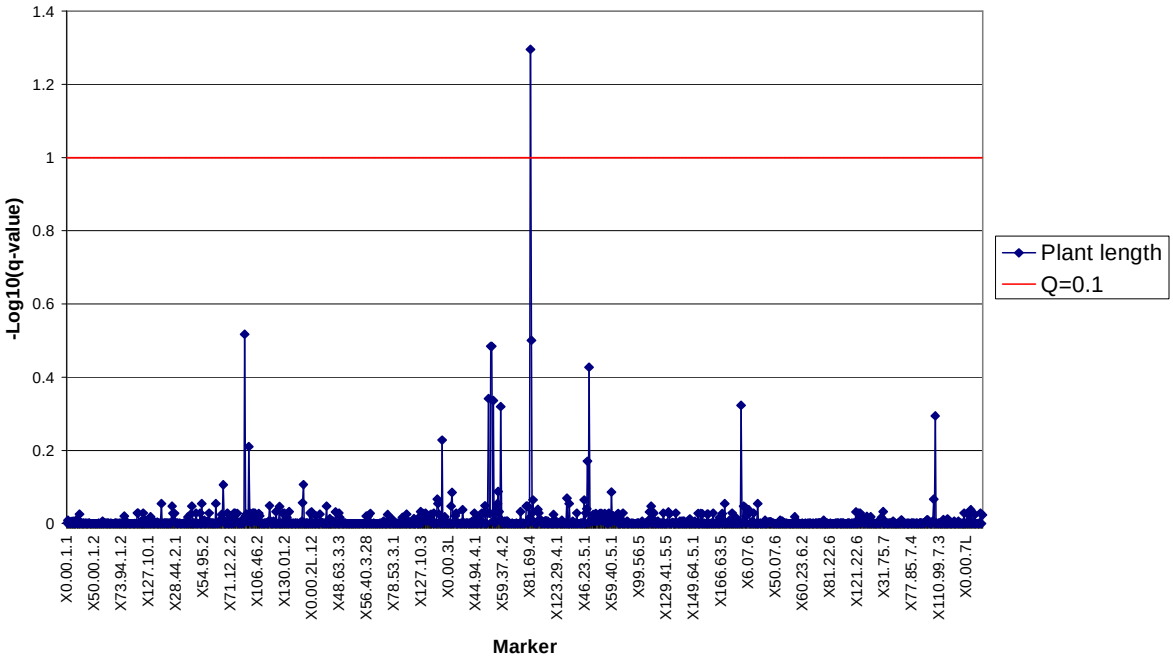


Figure 12-66. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

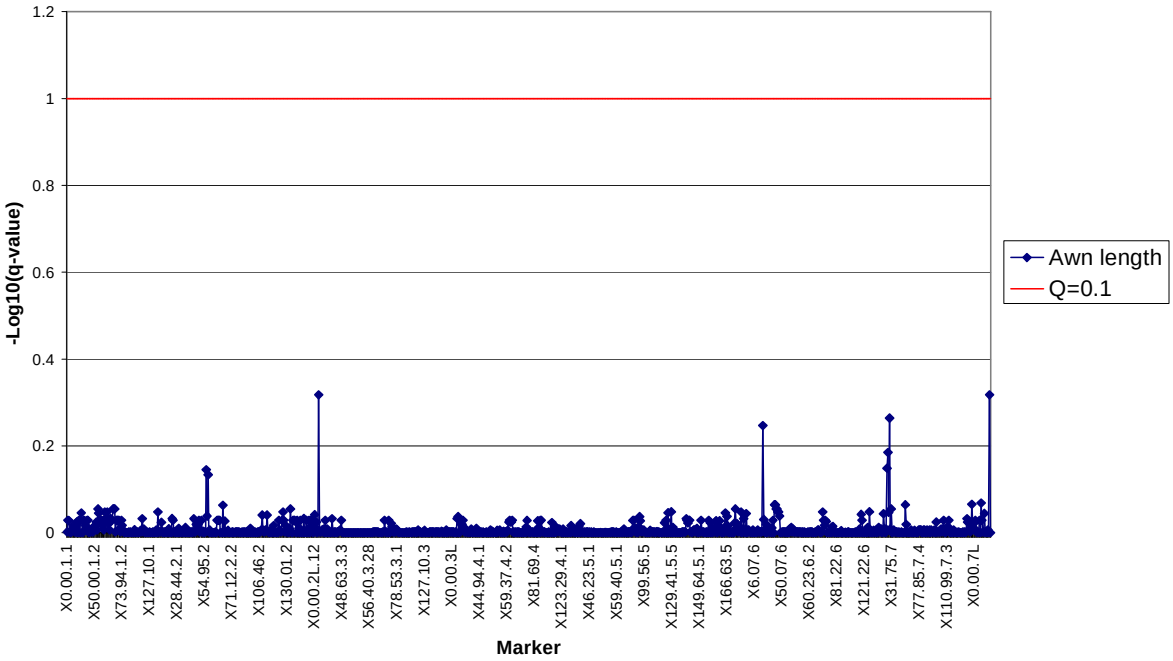


Figure 12-67. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

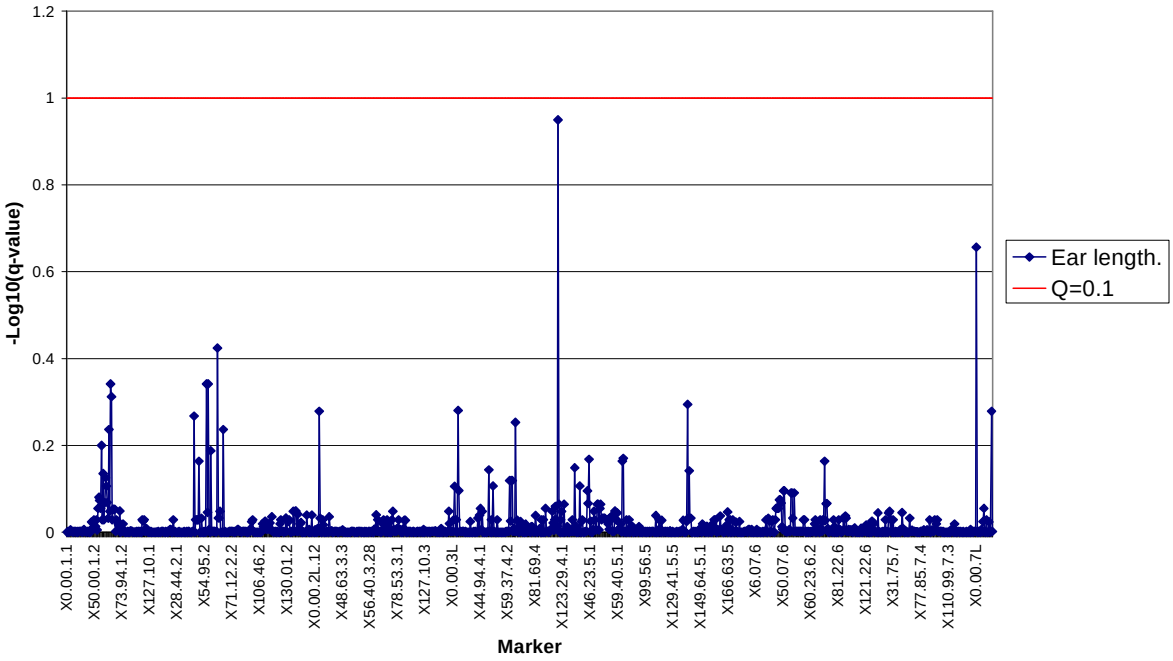


Figure 12-68. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

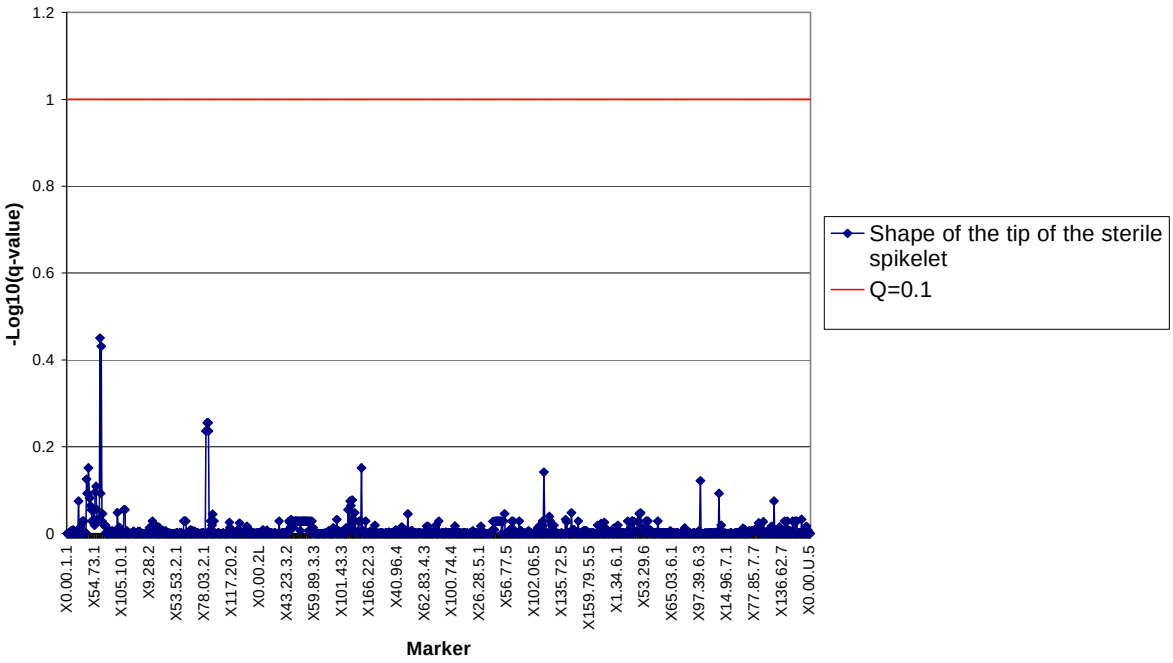


Figure 12-69. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

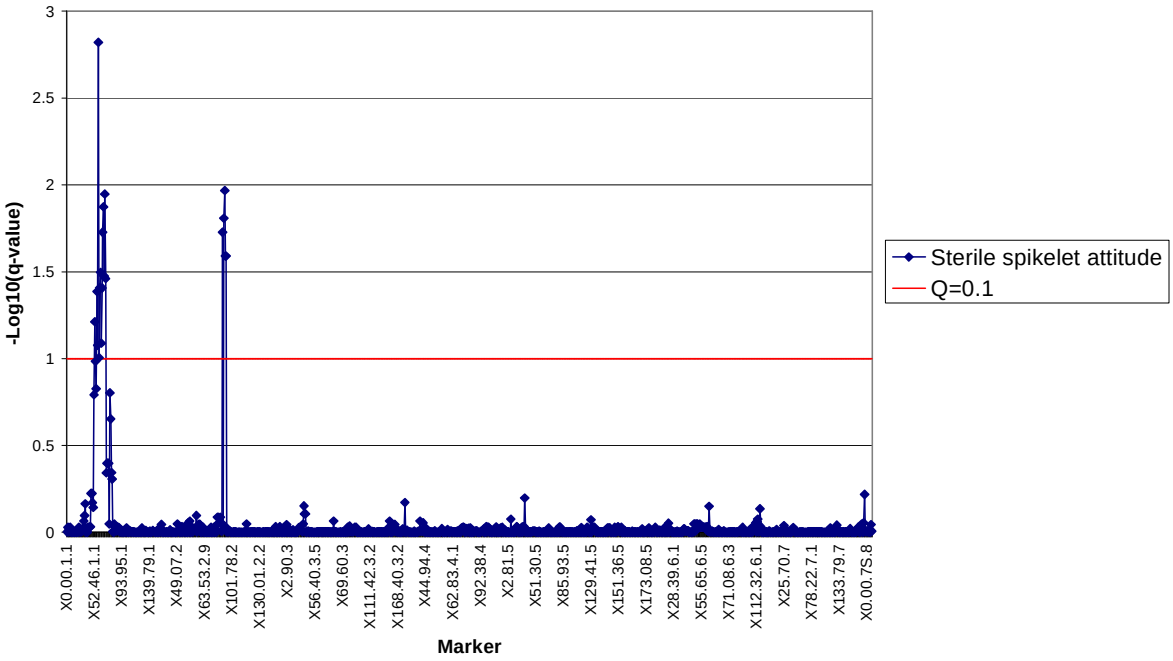


Figure 12-70. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

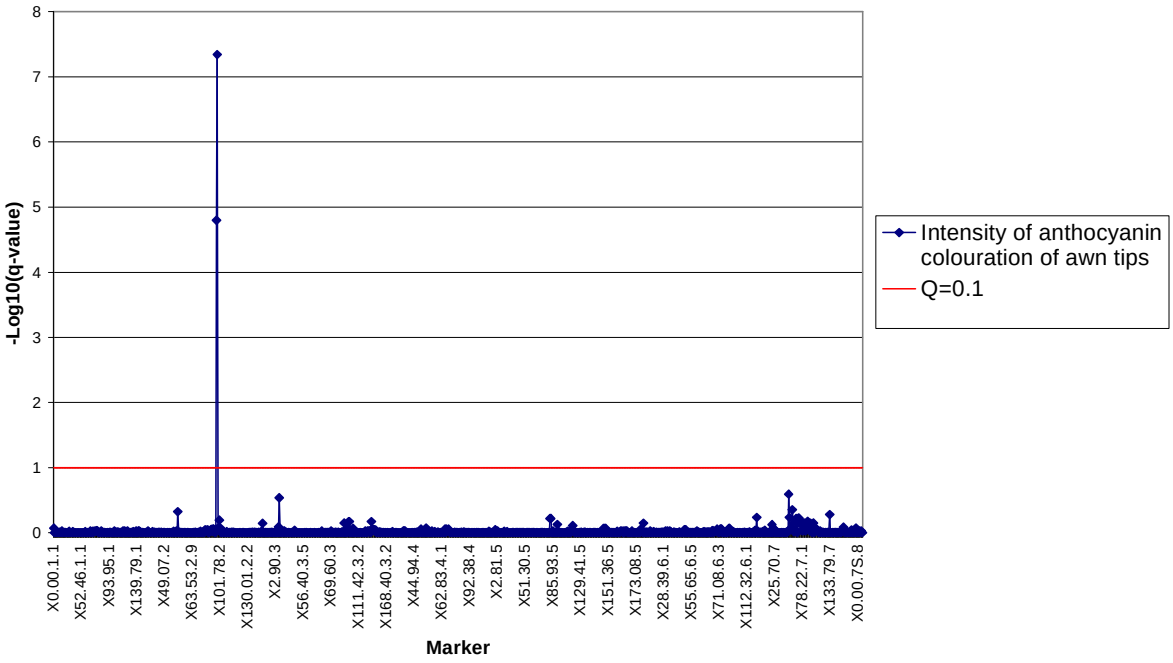


Figure 12-71. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

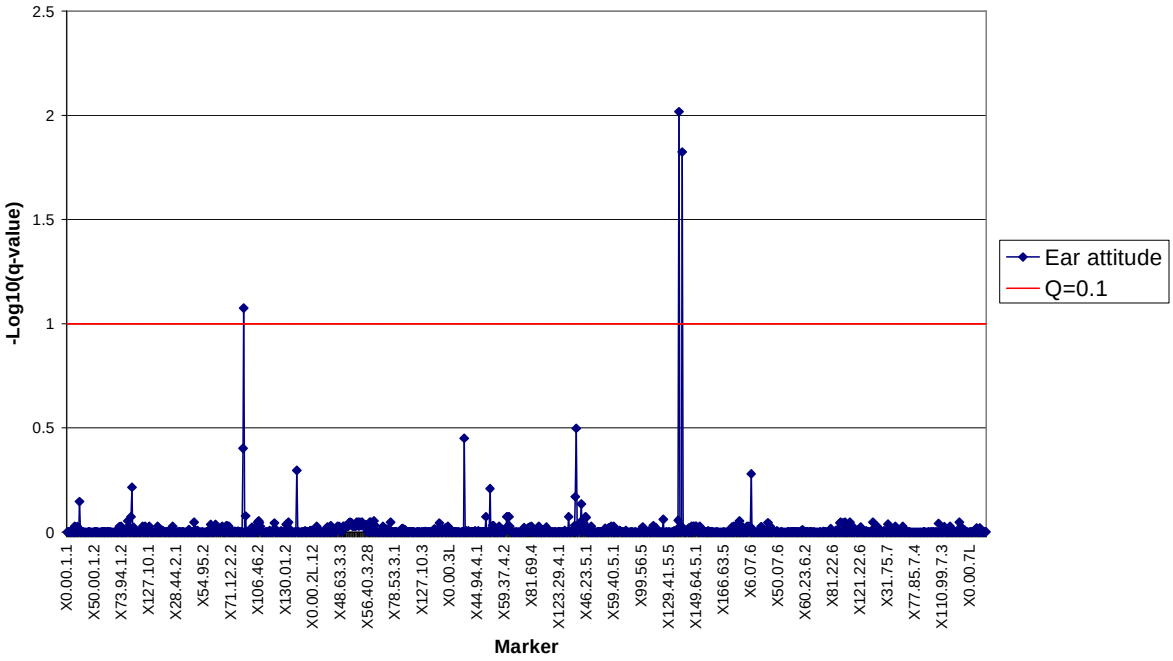


Figure 12-72. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

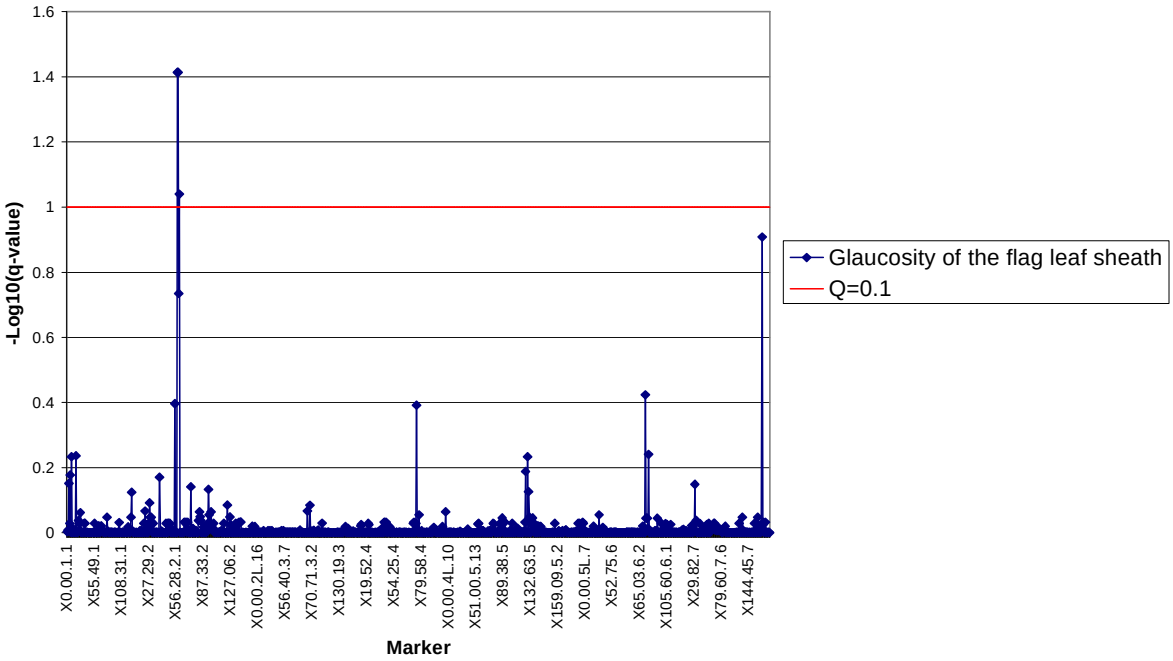


Figure 12-73. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

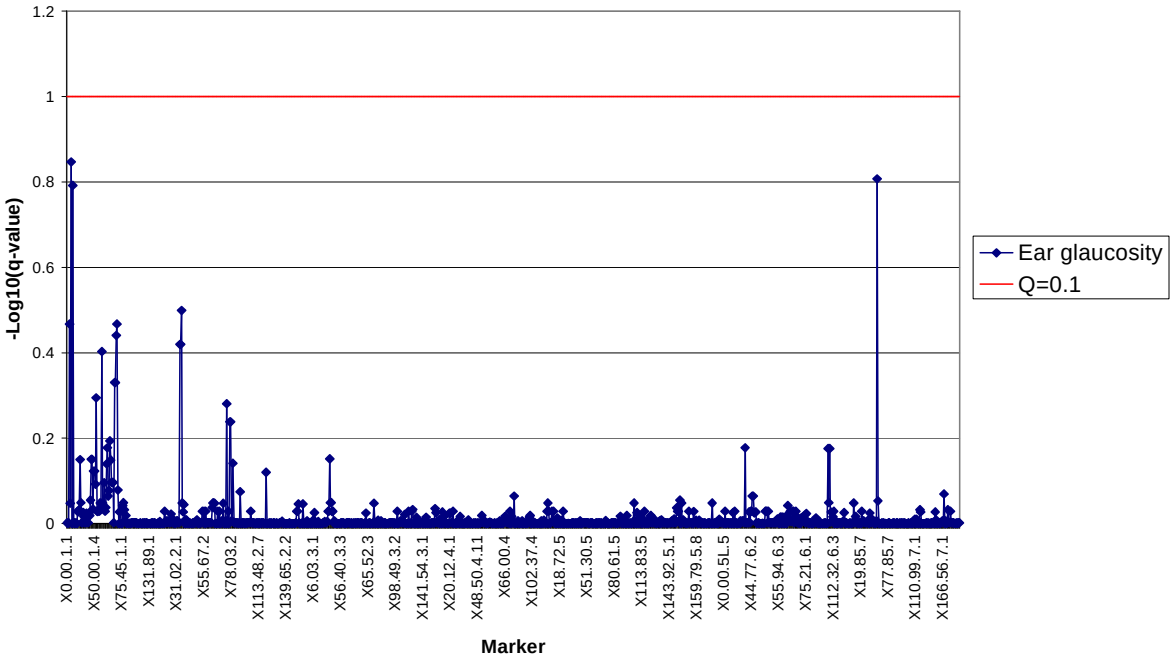


Figure 12-74. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

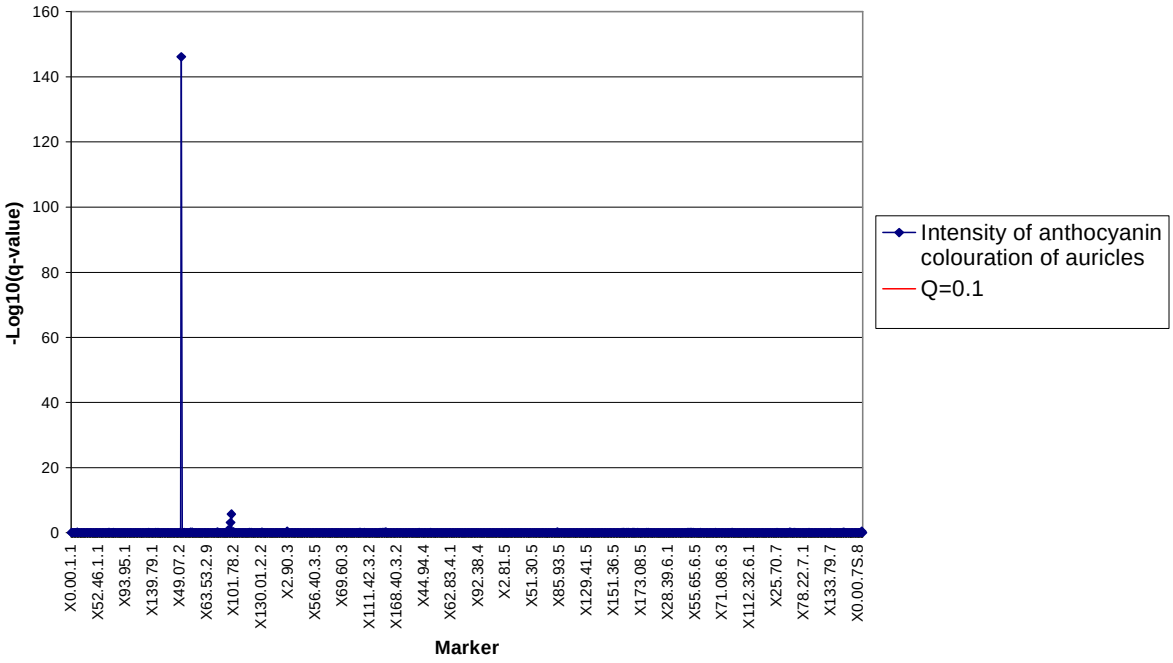


Figure 12-75. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

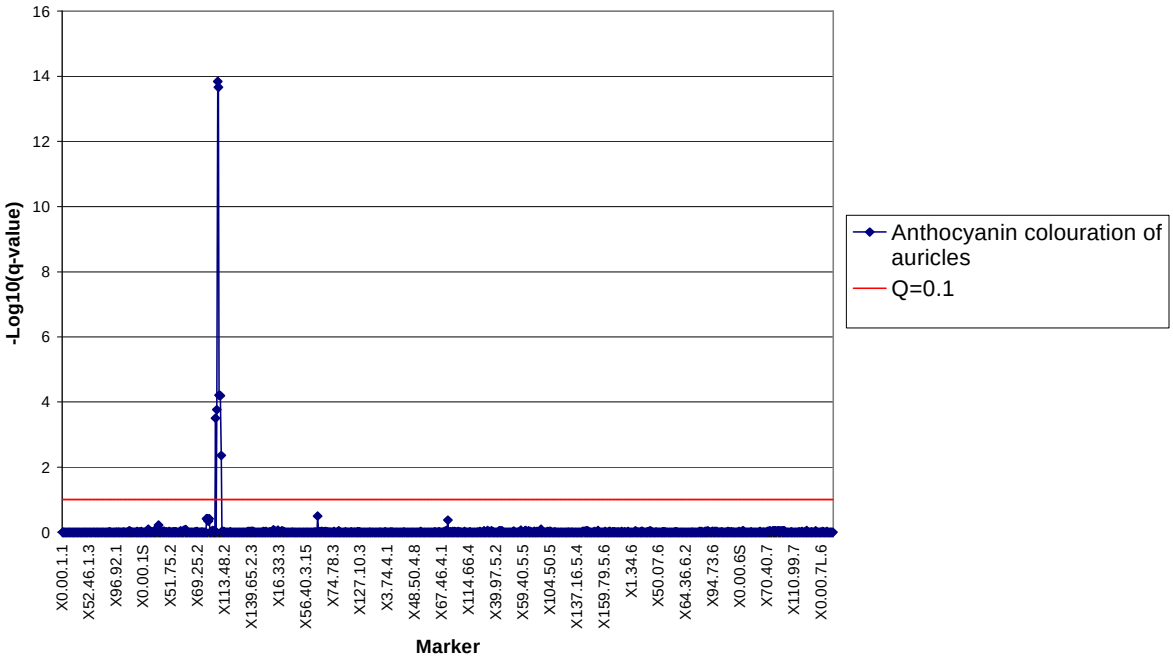


Figure 12-76. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

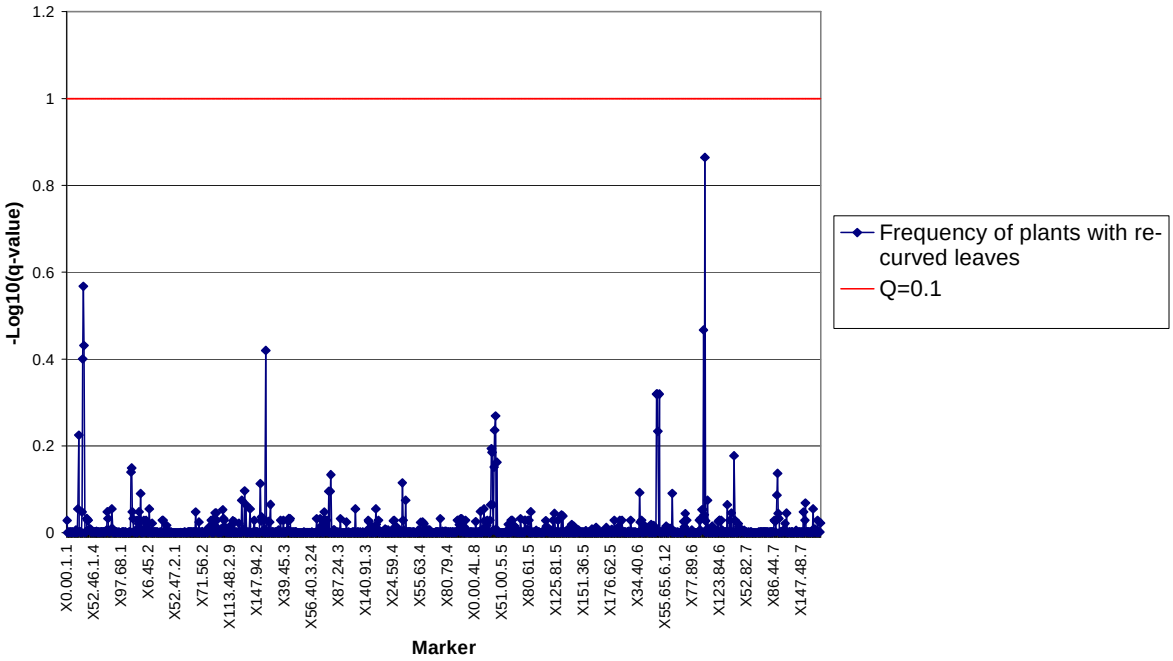


Figure 12-77. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

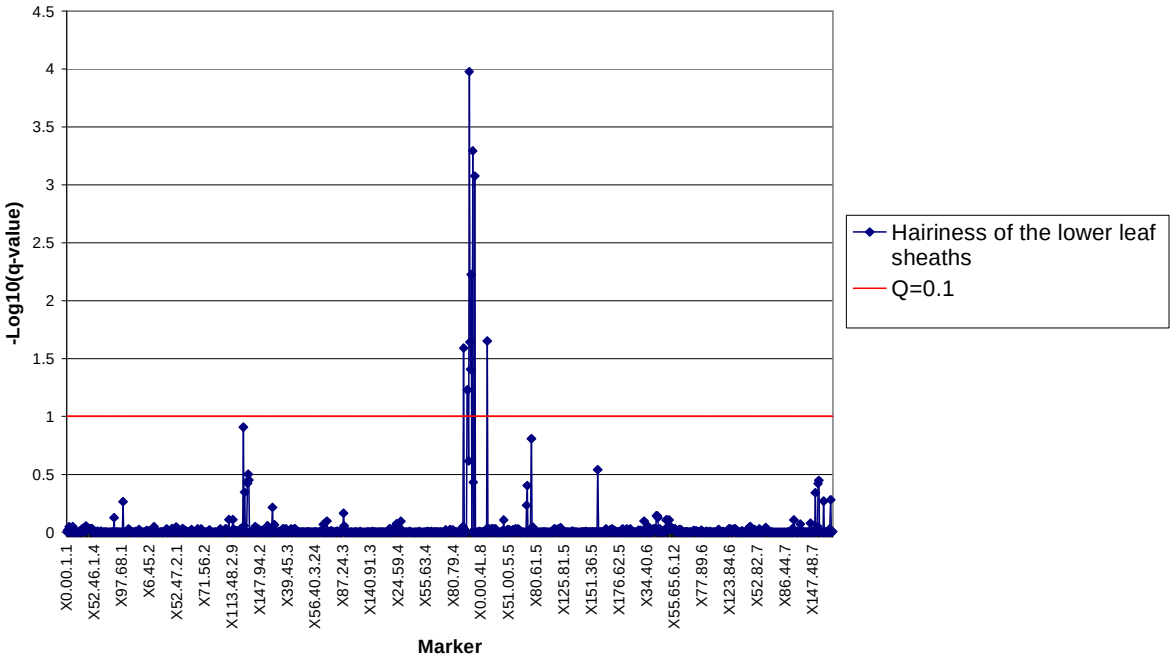


Figure 12-78. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

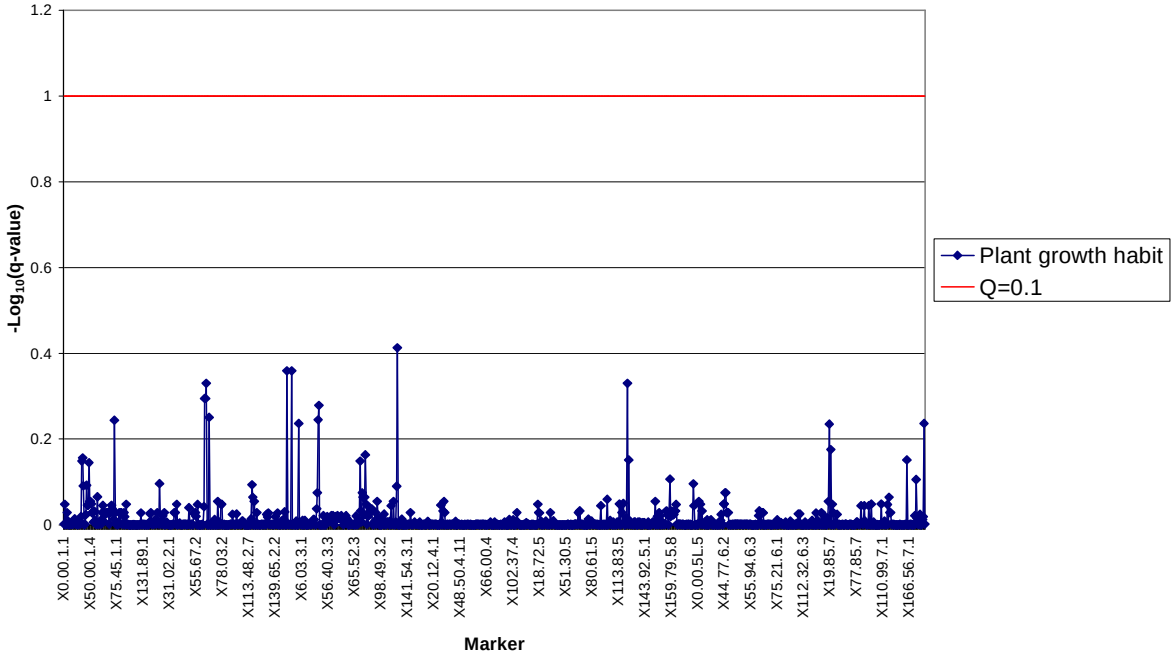


Figure 12-79. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

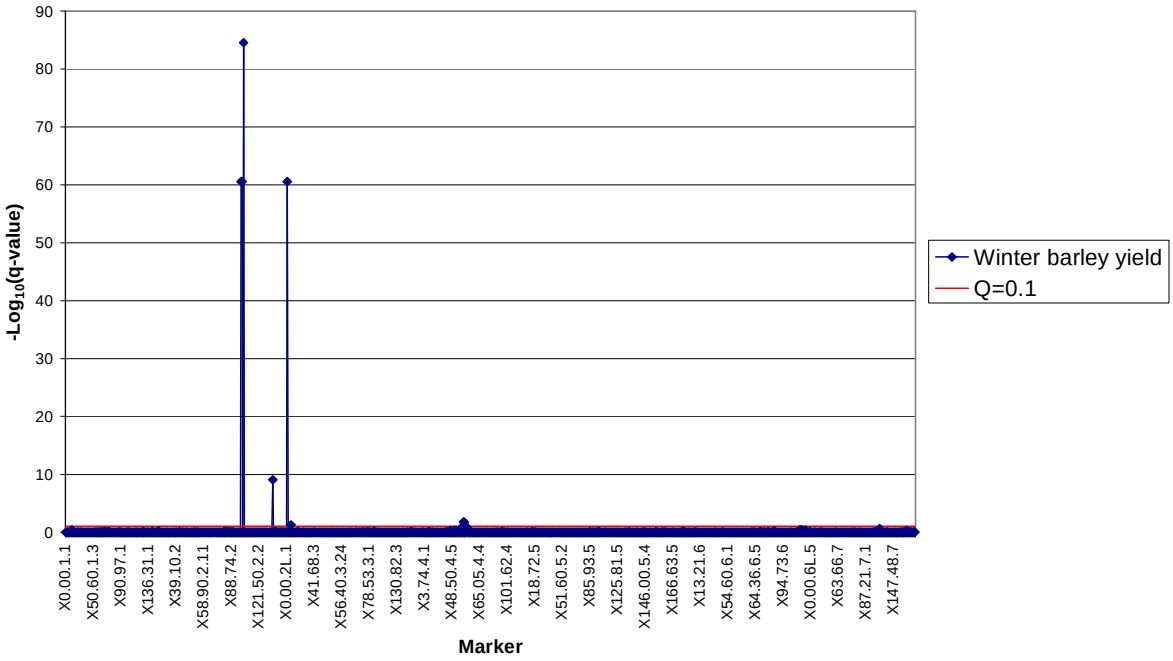


Figure 12-80. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

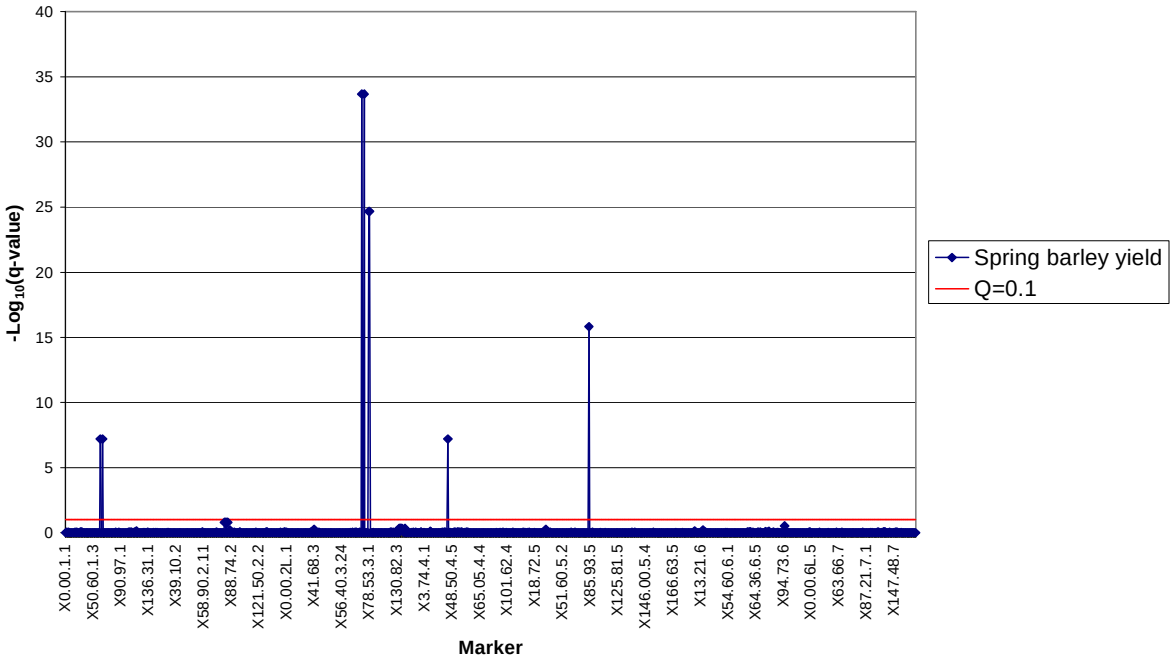


Figure 12-81. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

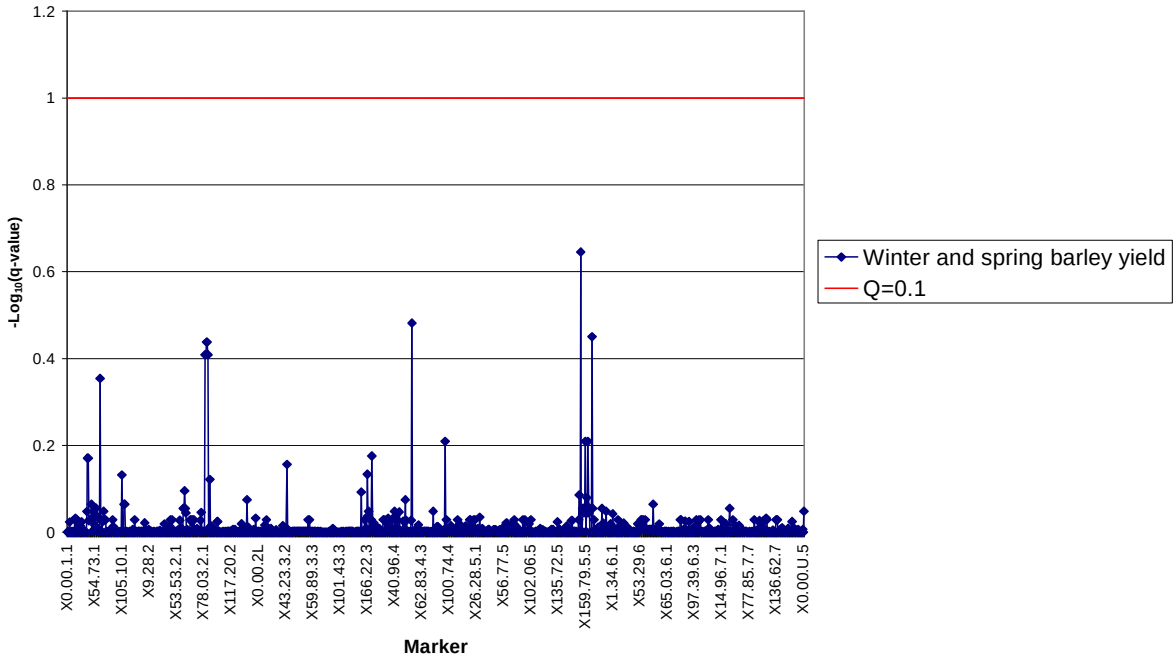


Figure 12-82. Association by SA + GC. Markers arranged in genetic map order. Trait as indicated in legend

12.9 Appendix 9. Association plots for barley phenotypes. Association estimated *with the effect of structure controlled using PCA +GC.*

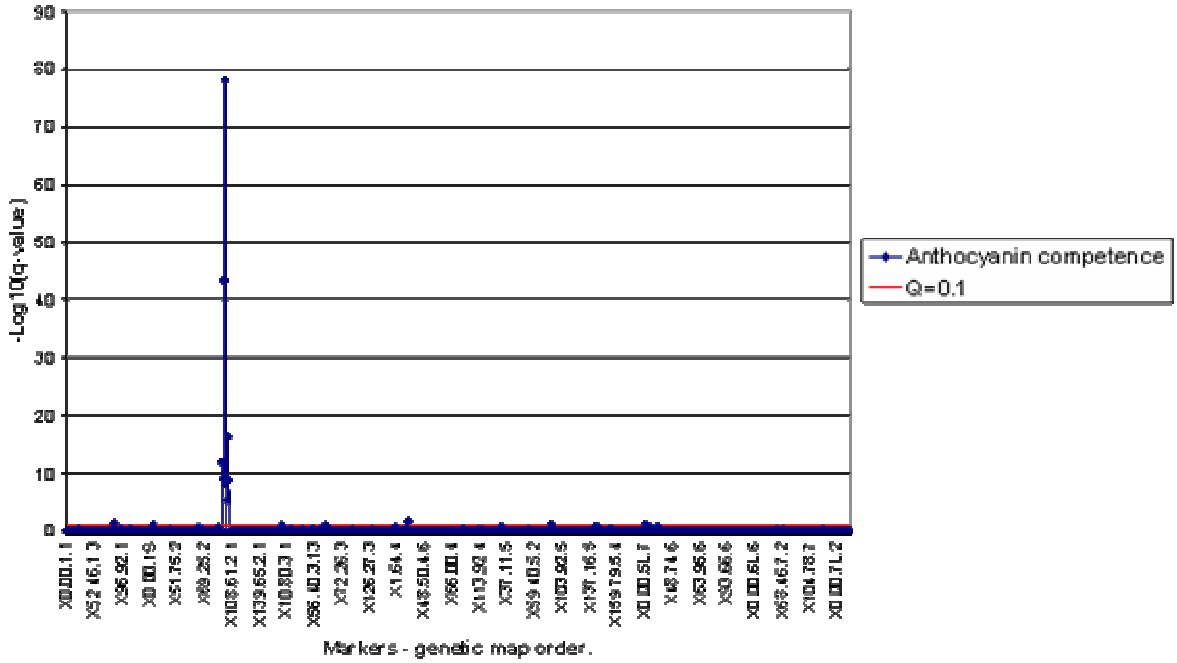


Figure 12-83. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

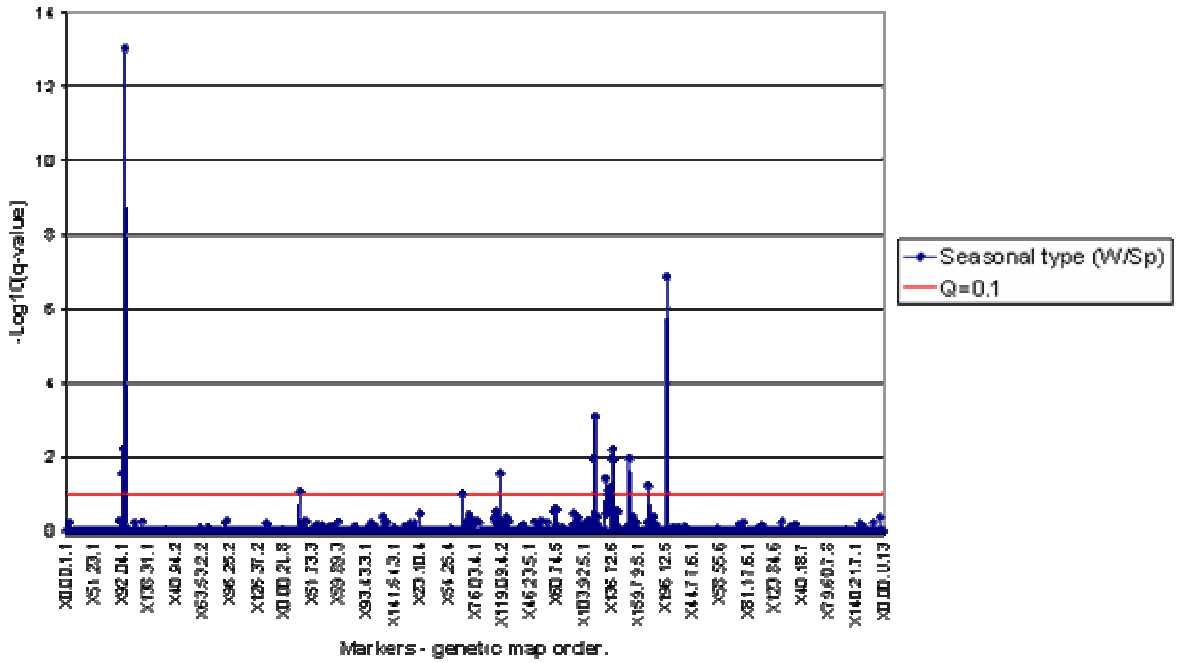


Figure 12-84. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

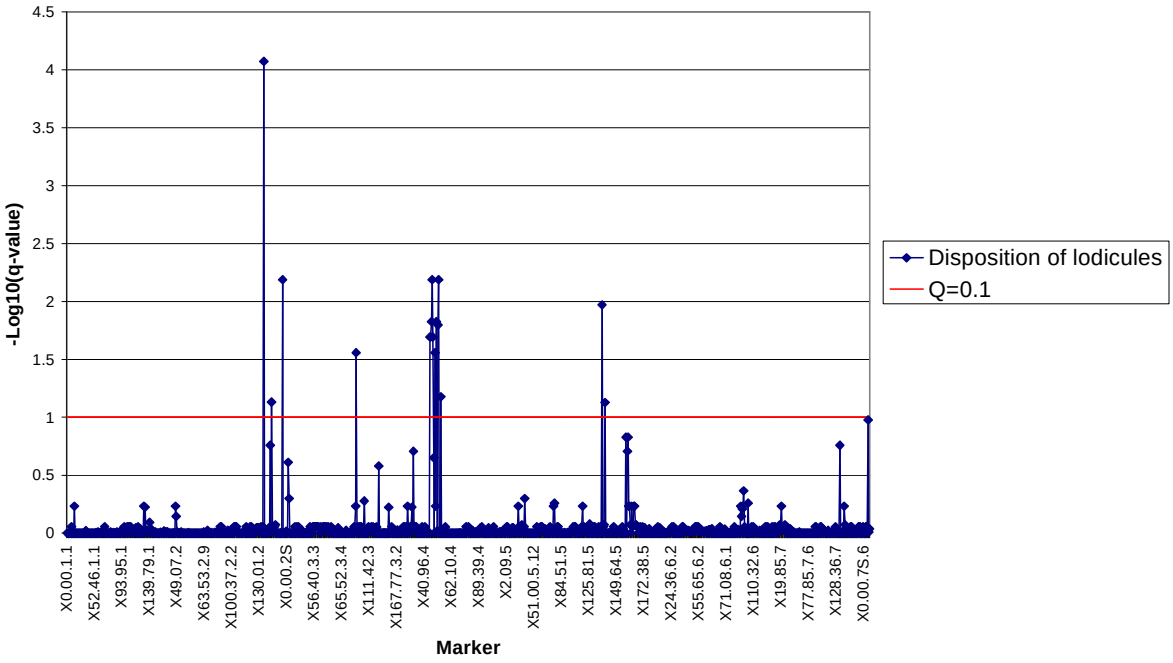


Figure 12-85. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

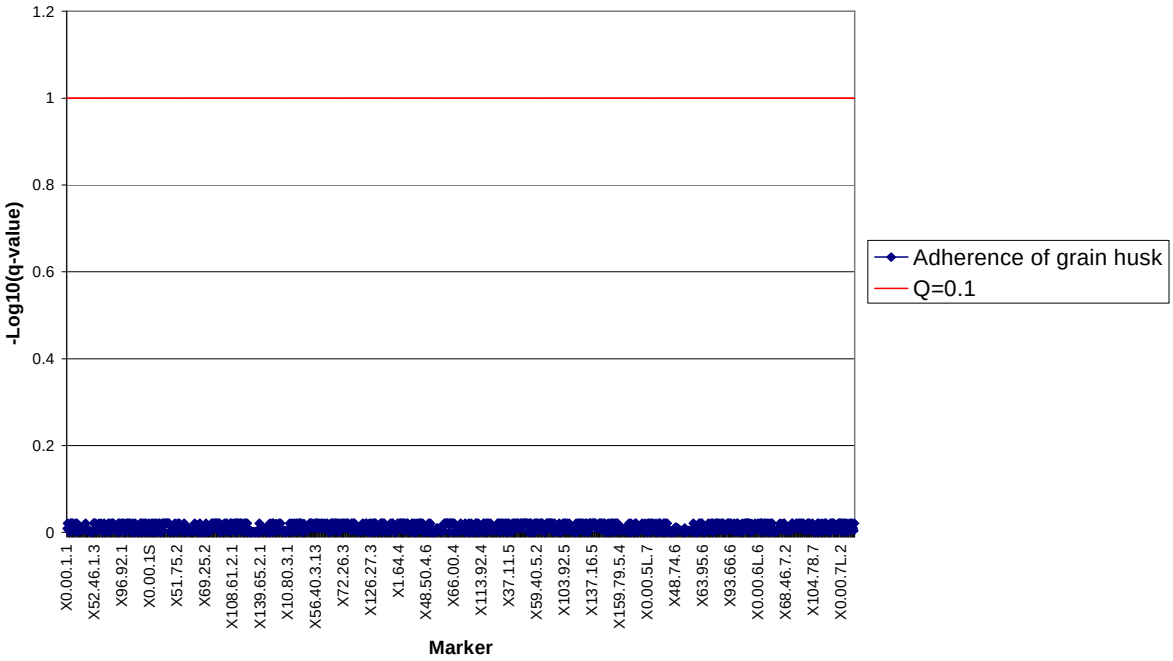


Figure 12-86. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

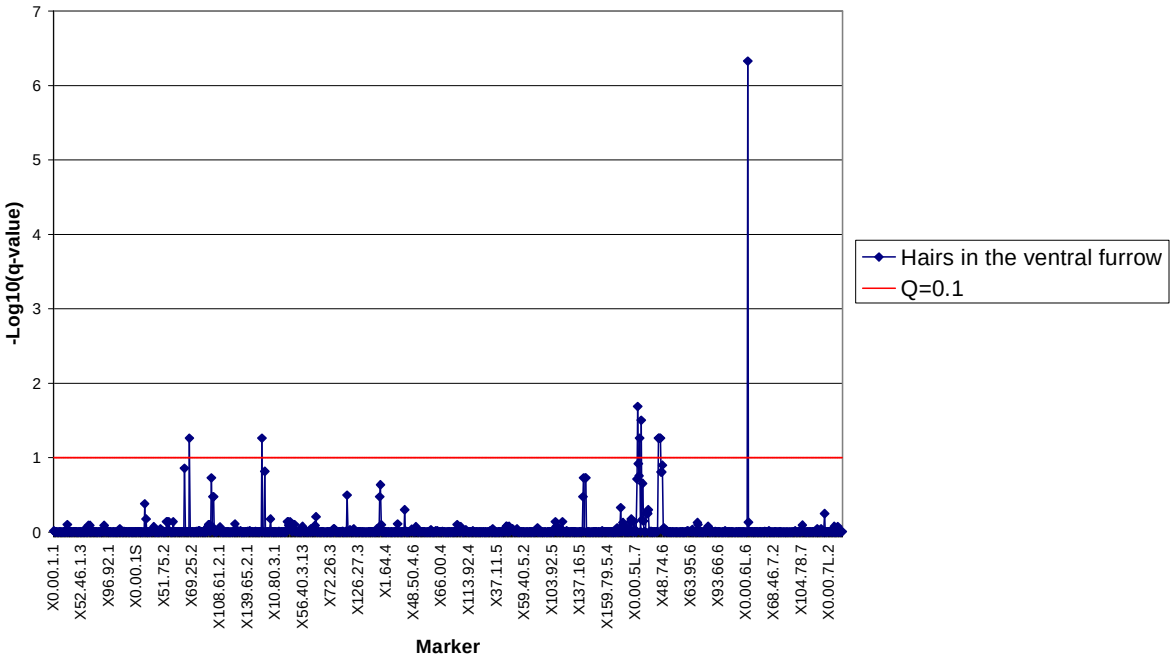


Figure 12-87. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

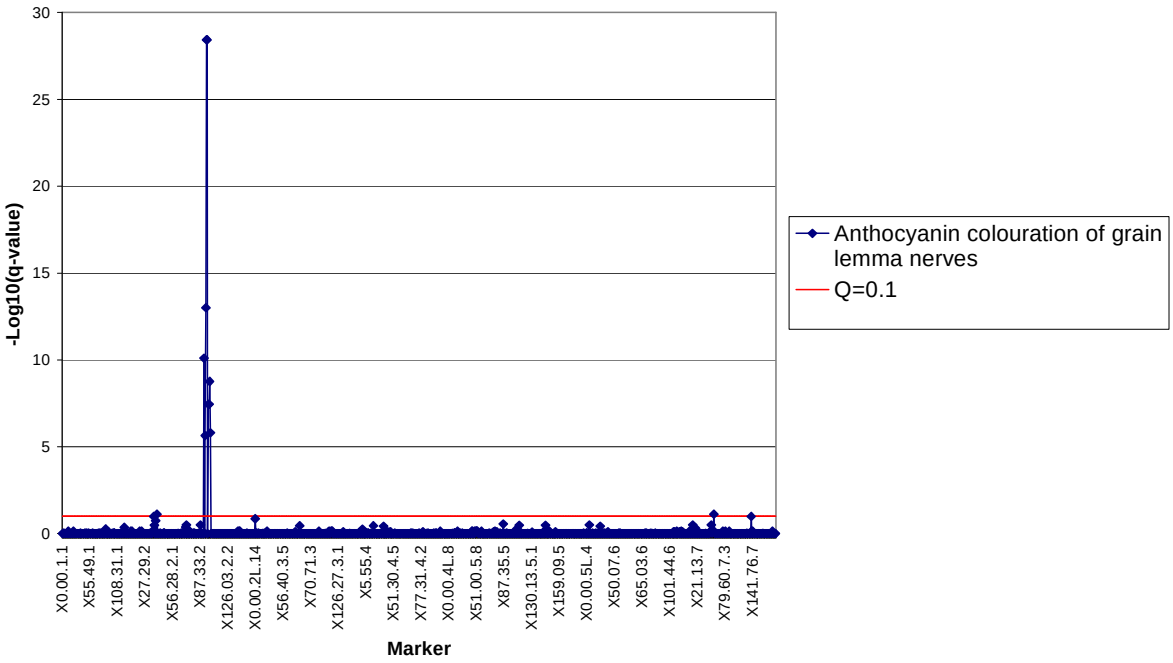


Figure 12-88. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

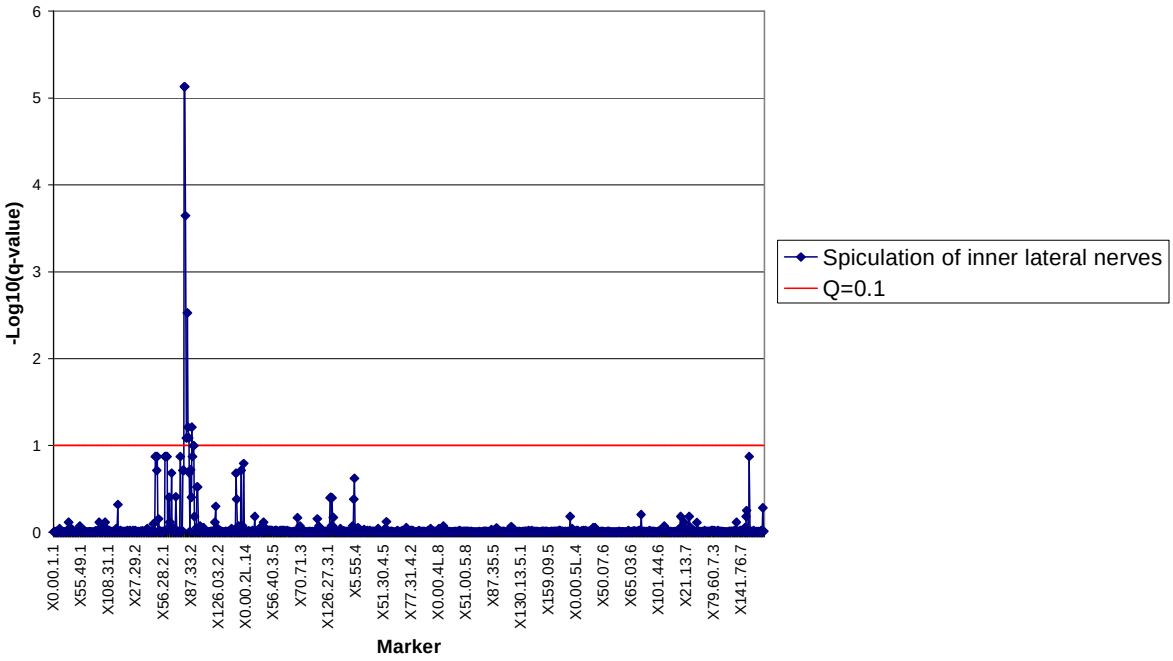


Figure 12-89. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

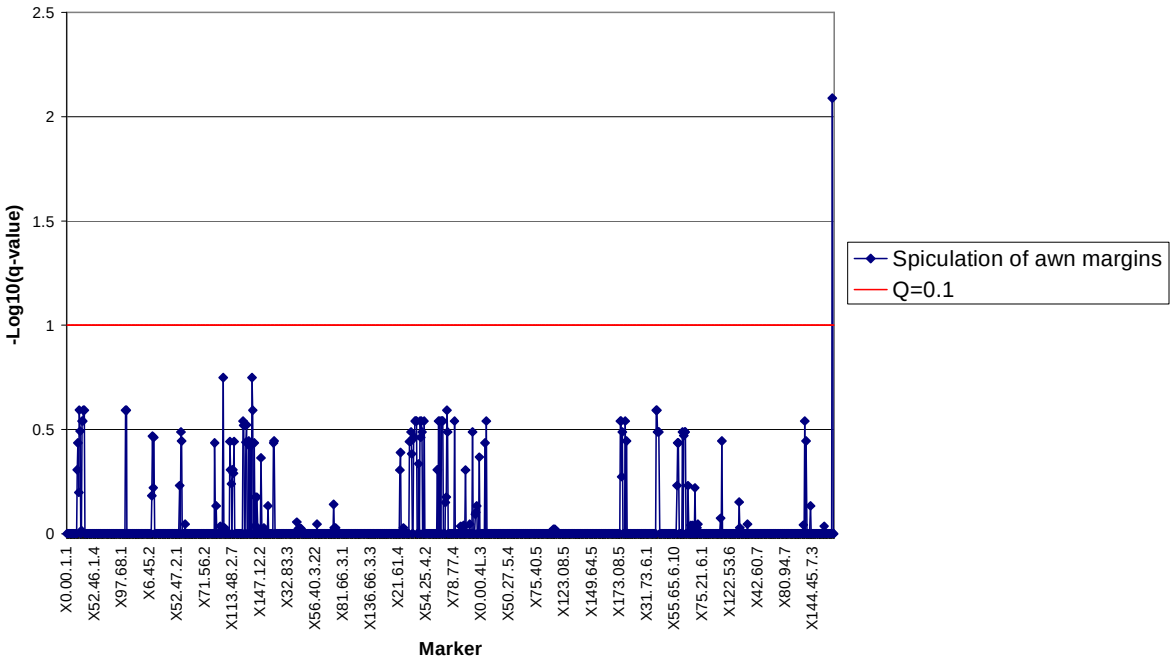


Figure 12-90. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

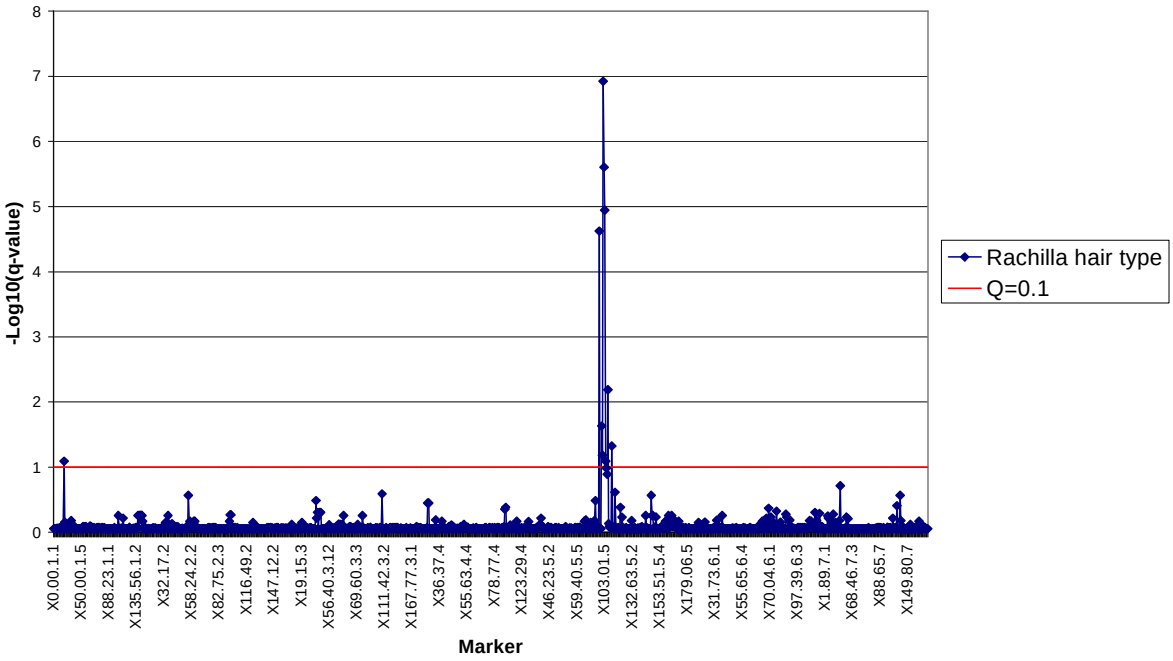


Figure 12-91. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

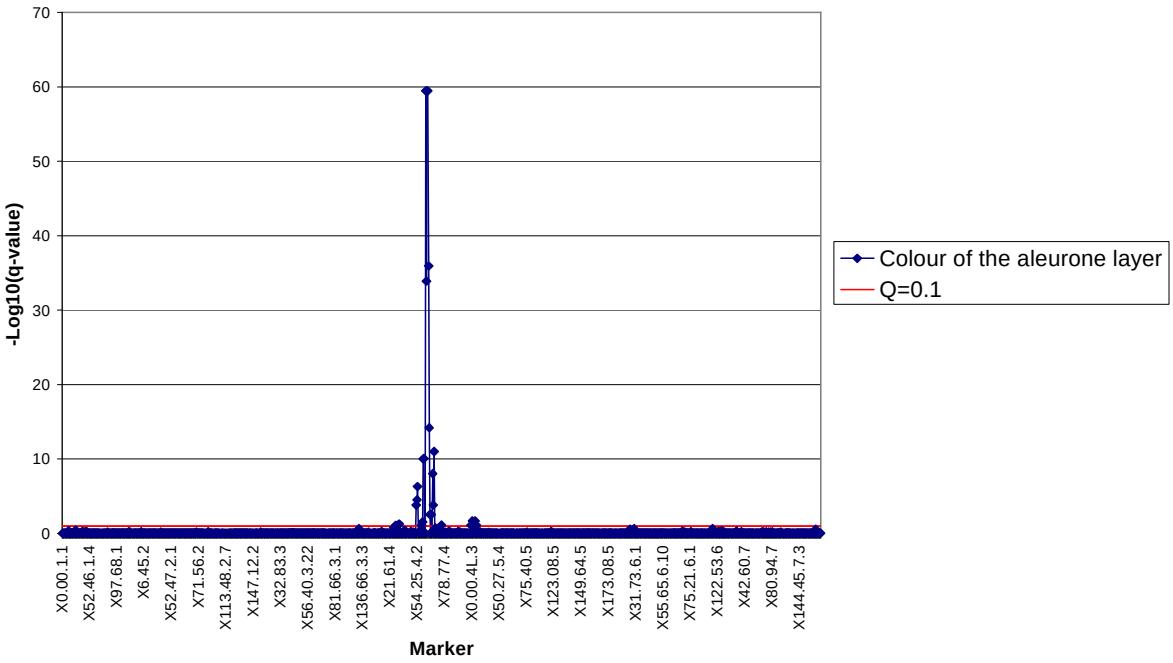


Figure 12-92. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

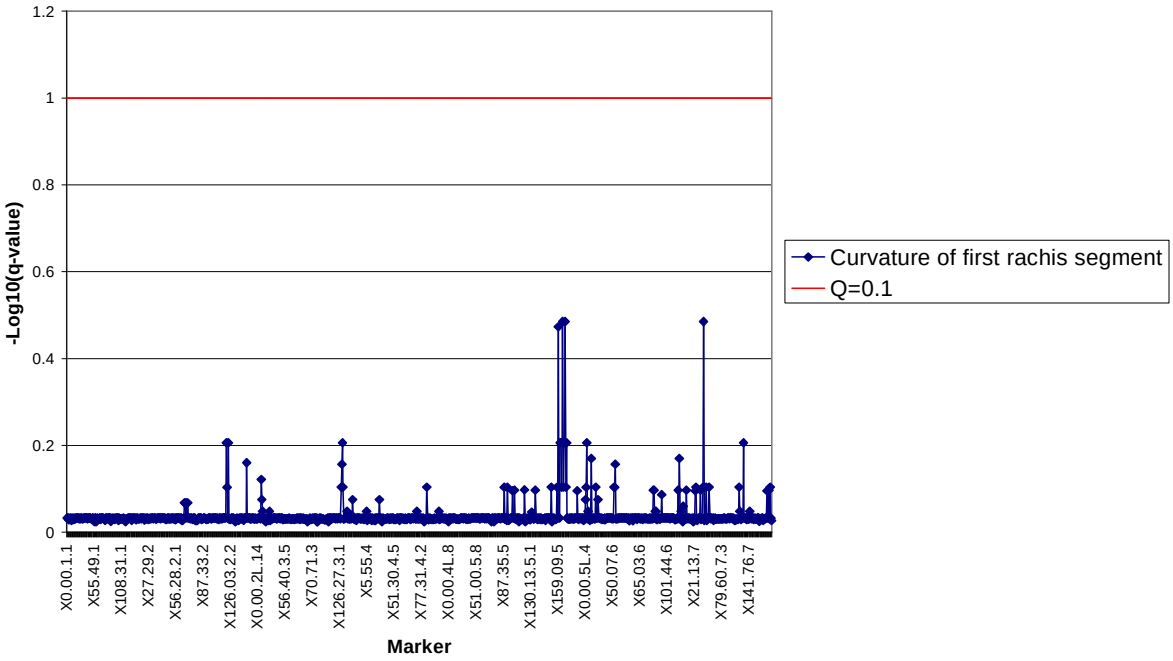


Figure 12-93. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

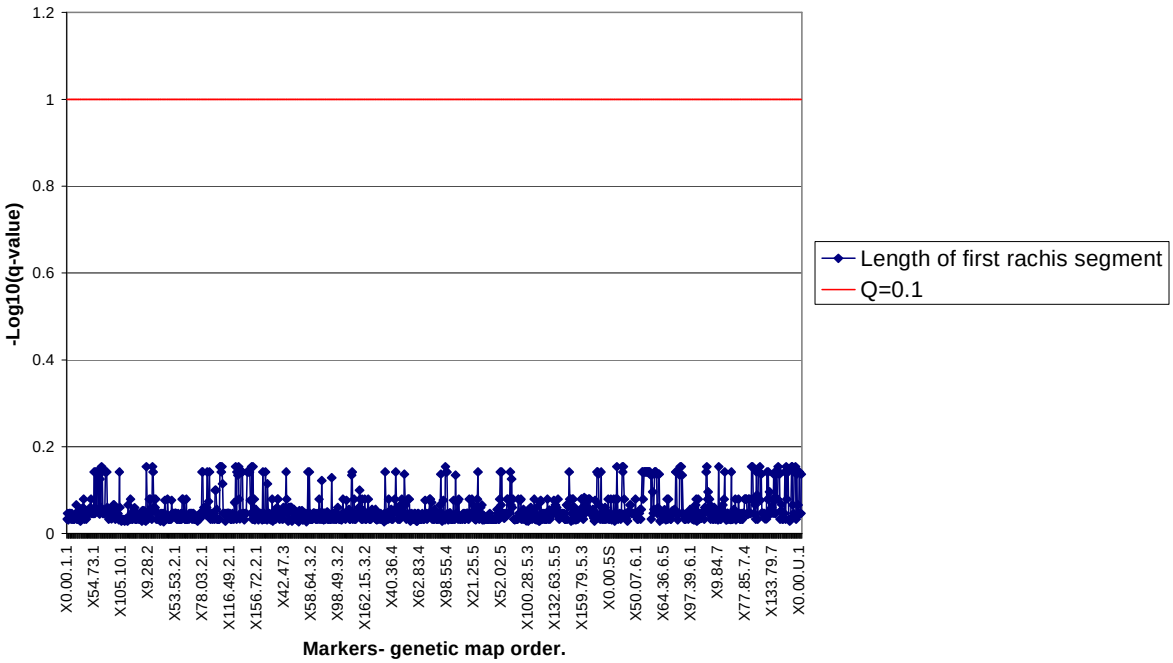


Figure 12-94. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

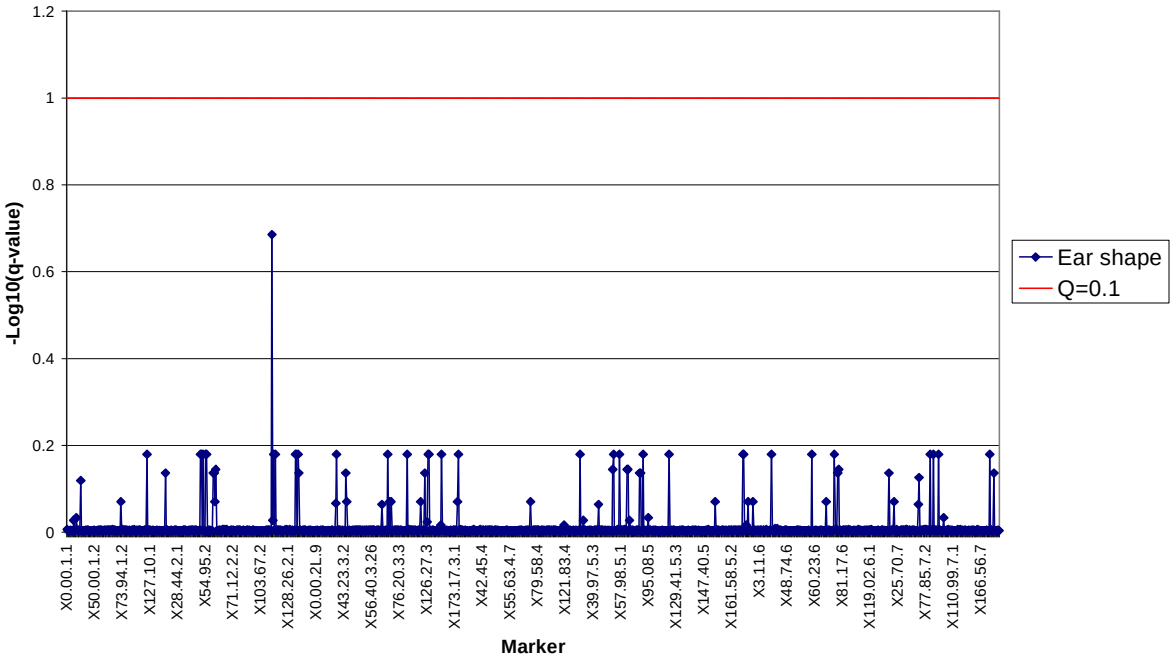


Figure 12-95. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

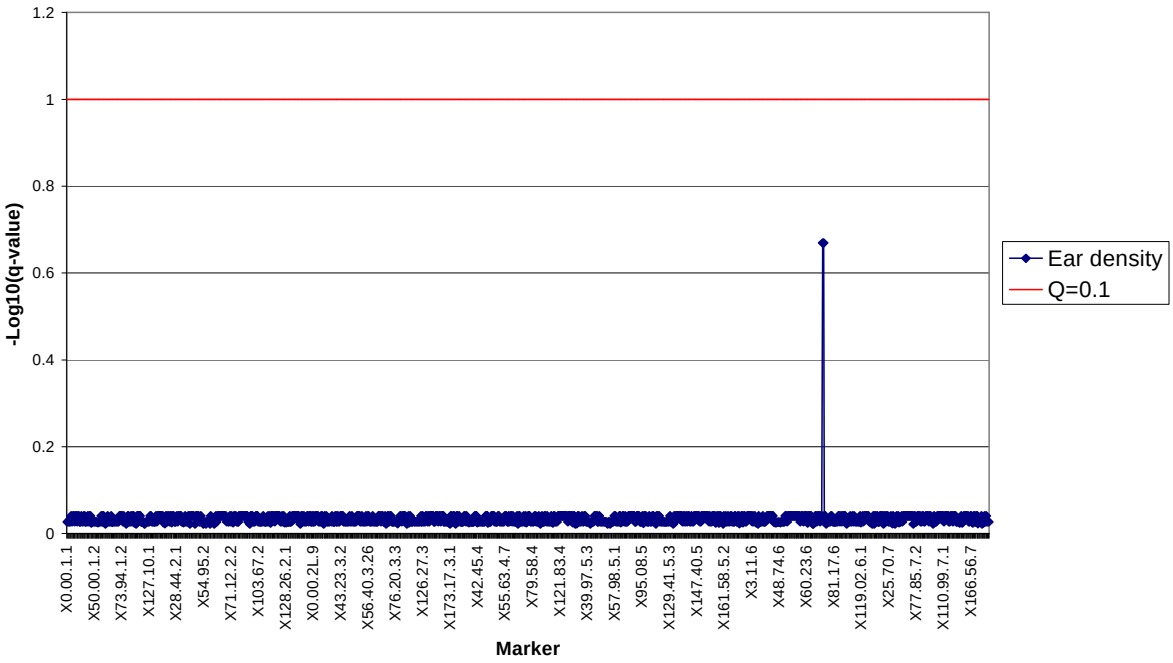


Figure 12-96. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

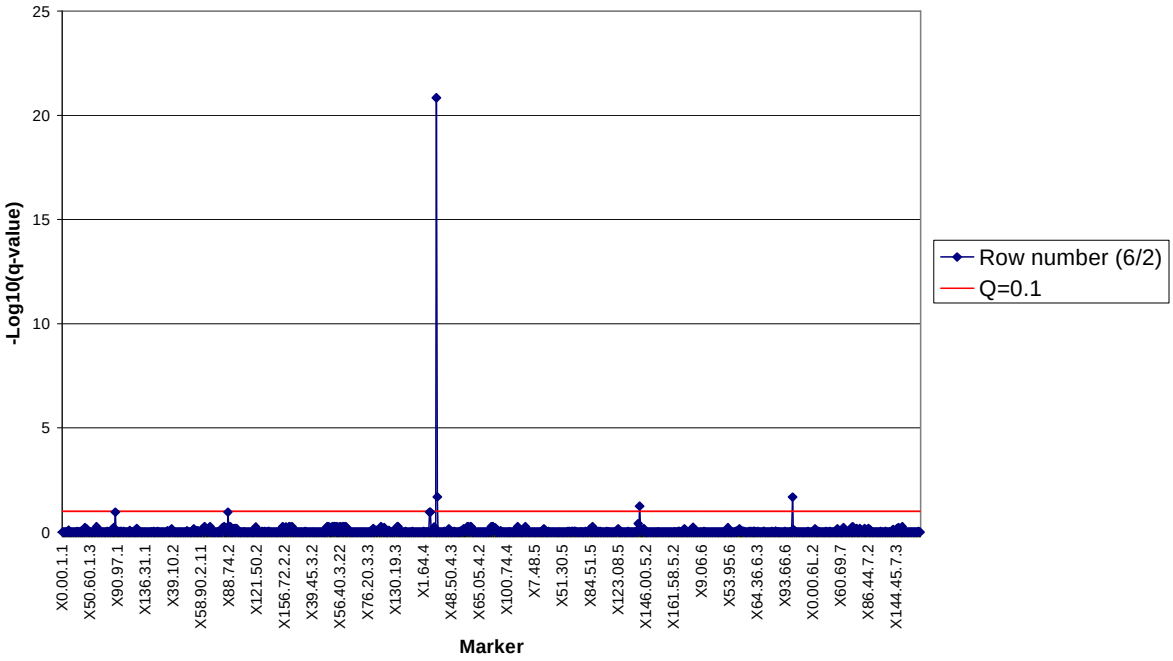


Figure 12-97. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

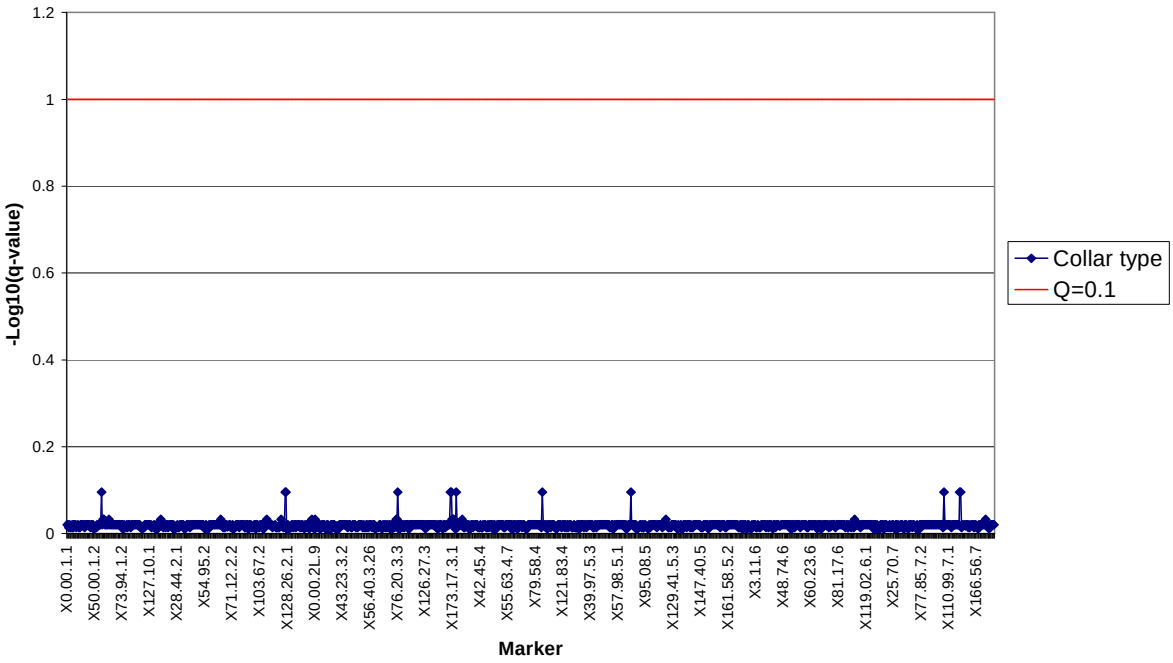


Figure 12-98. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

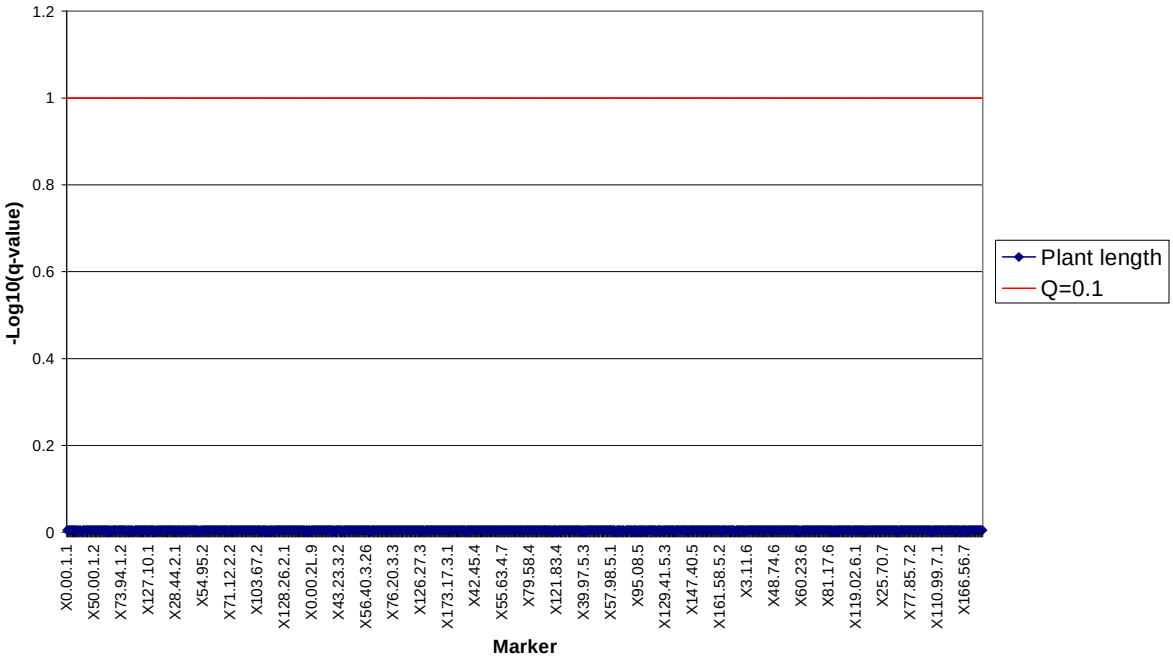


Figure 12-99. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

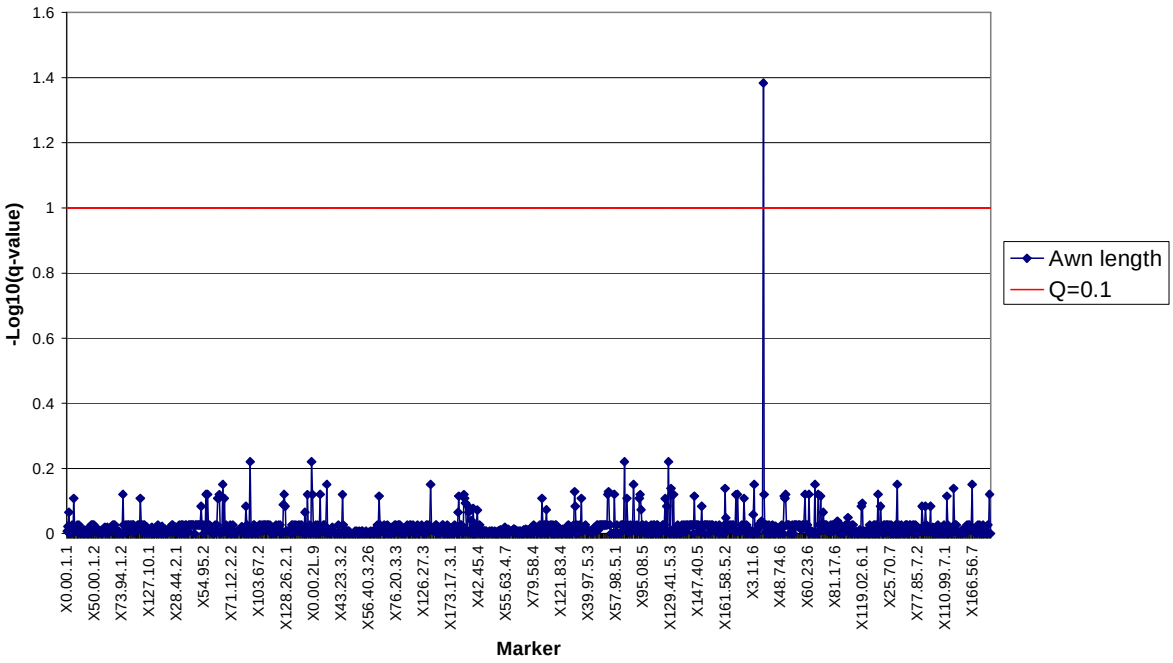


Figure 12-100. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

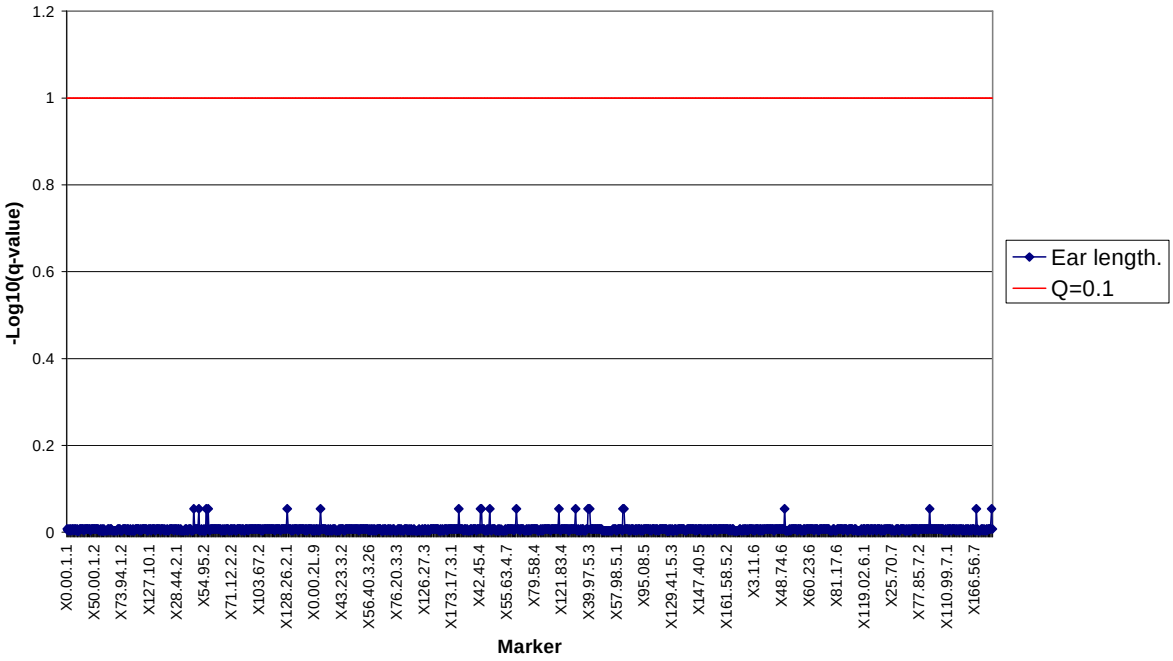


Figure 12-101. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

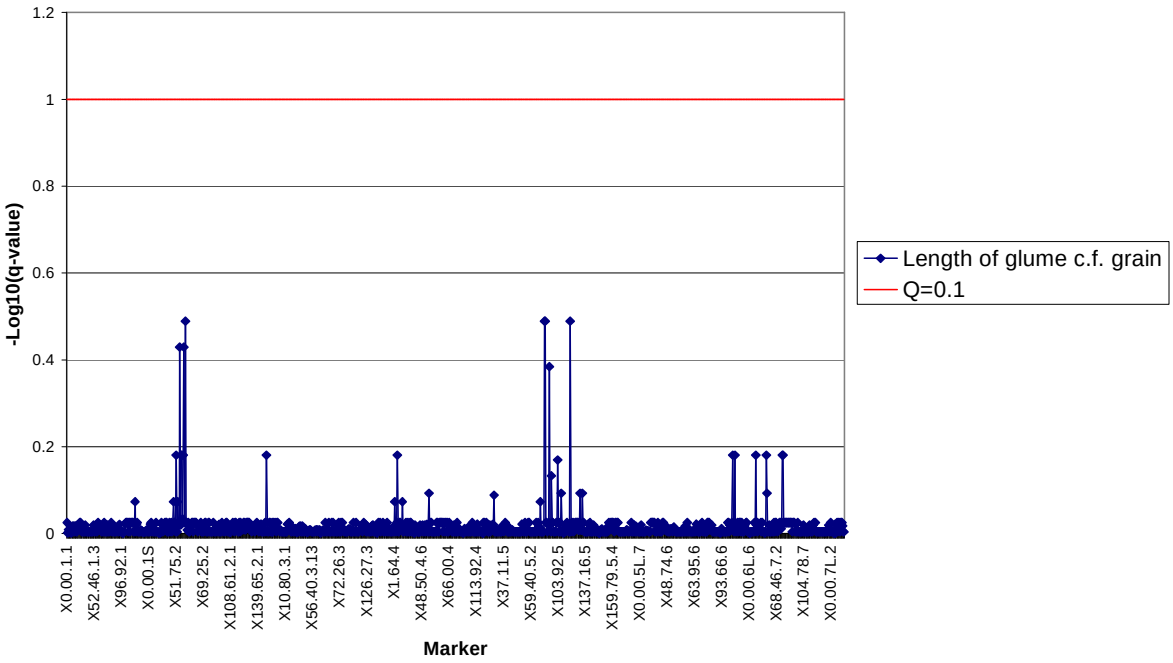


Figure 12-102. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

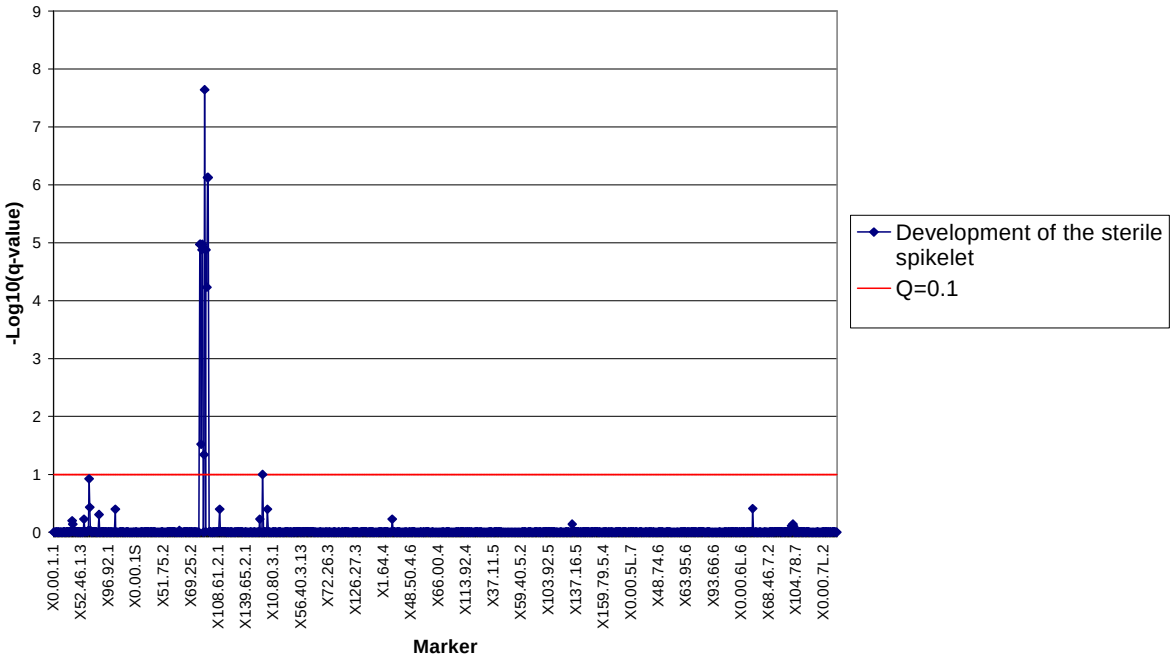


Figure 12-103. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

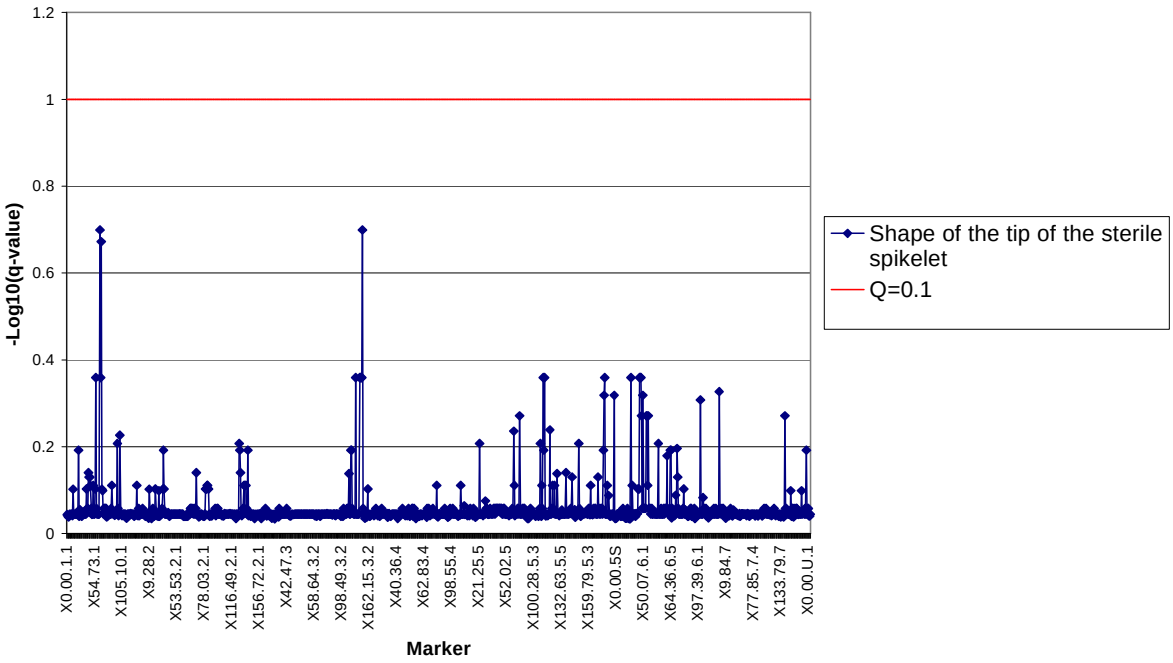


Figure 12-104. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

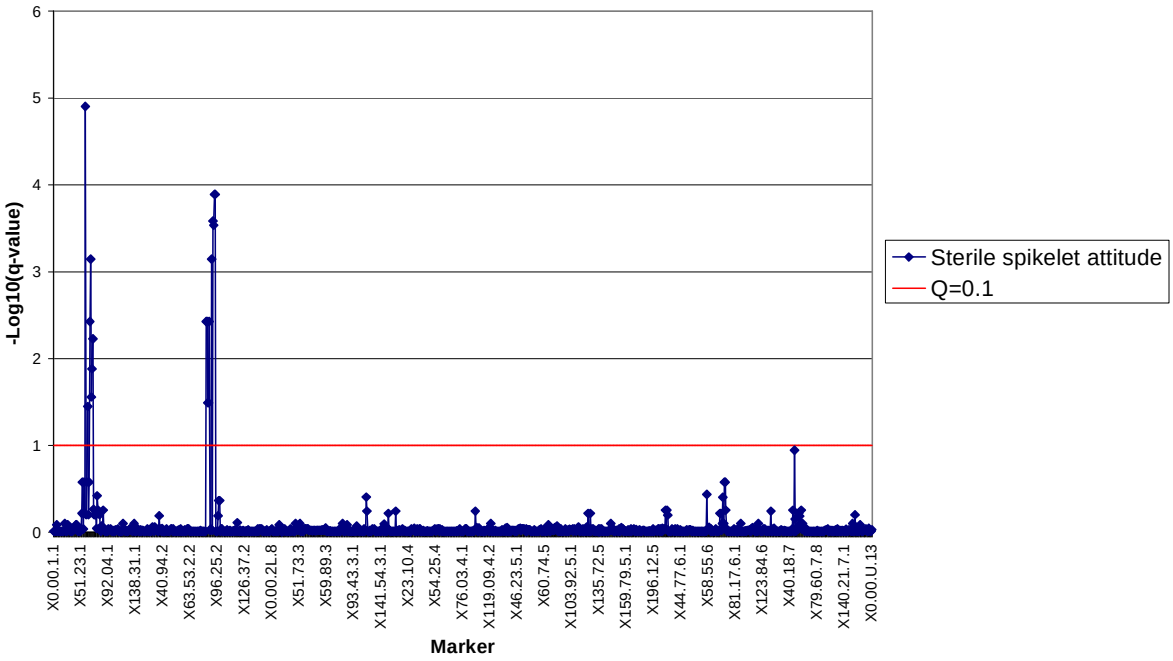


Figure 12-105. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

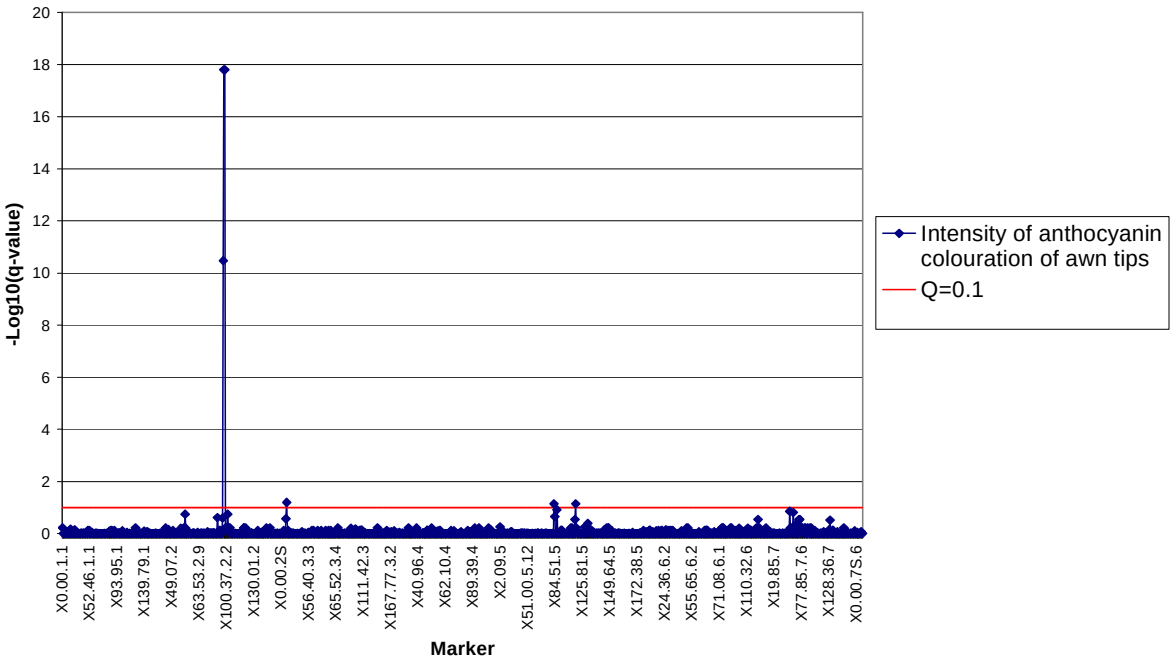


Figure 12-106. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

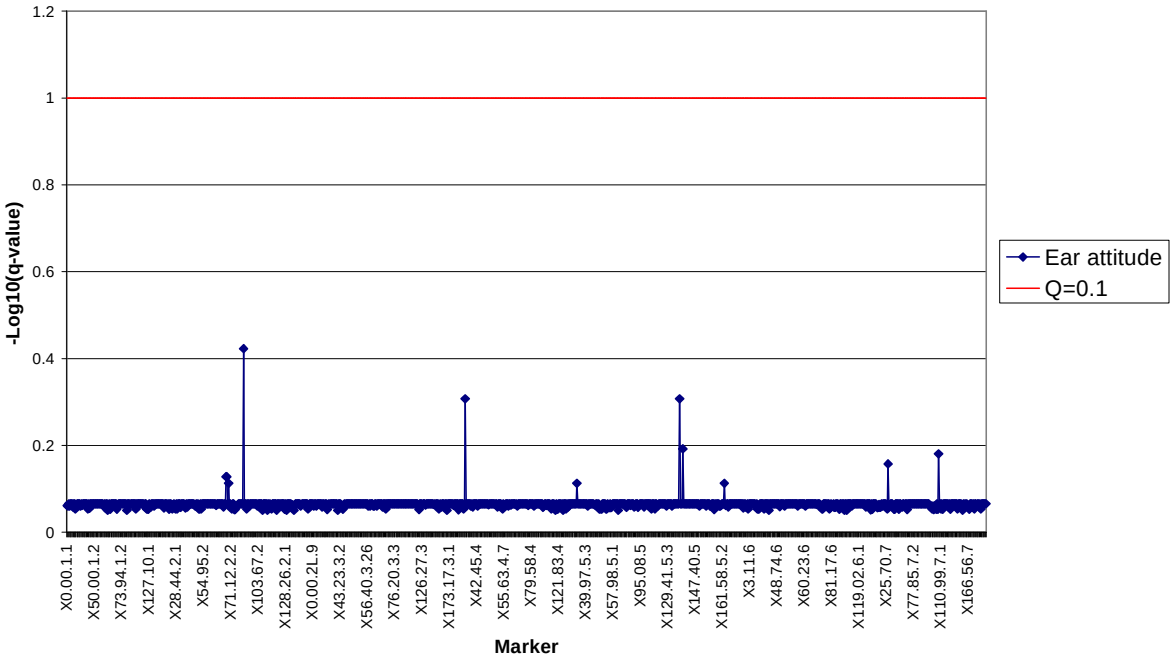


Figure 12-107. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

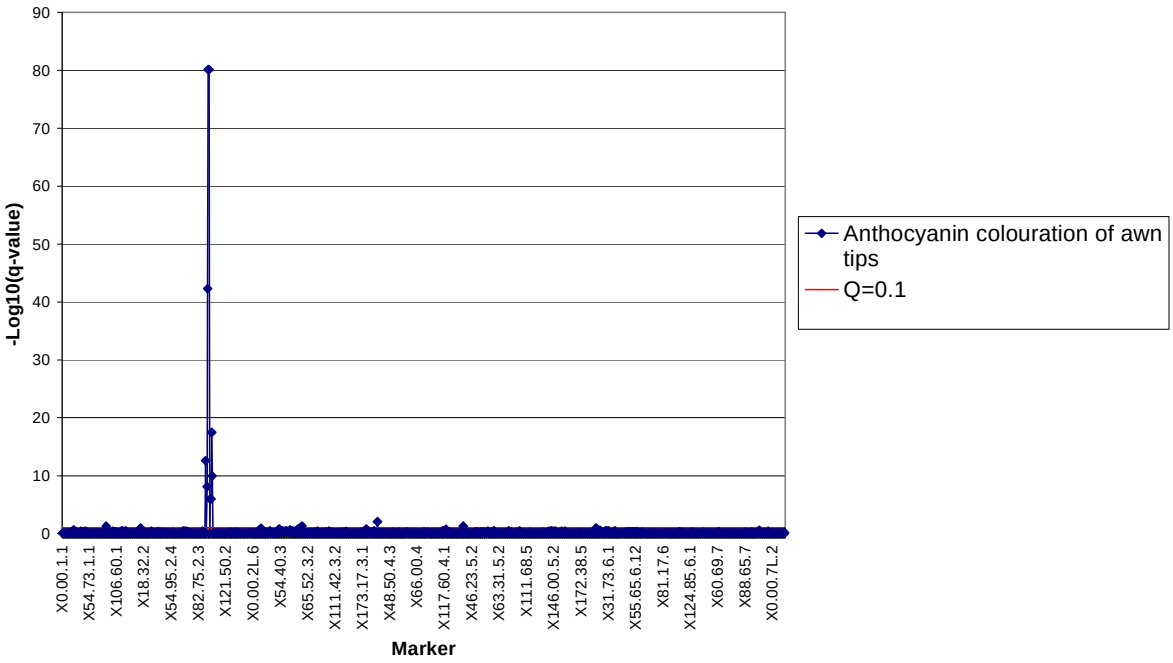


Figure 12-108. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

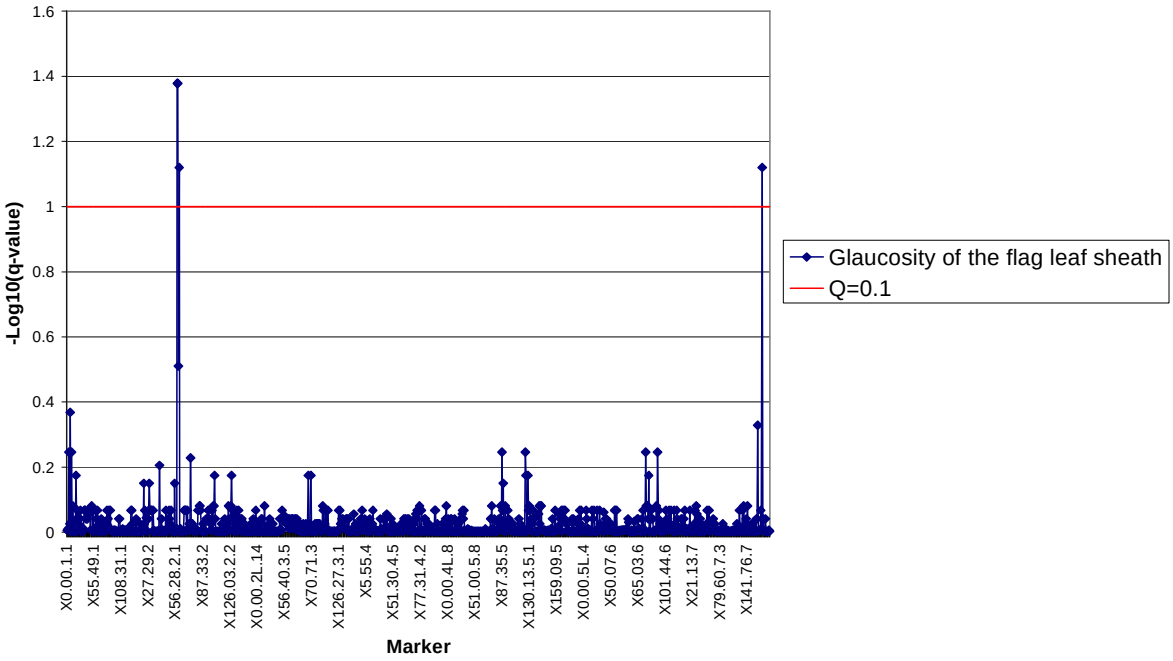


Figure 12-109. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

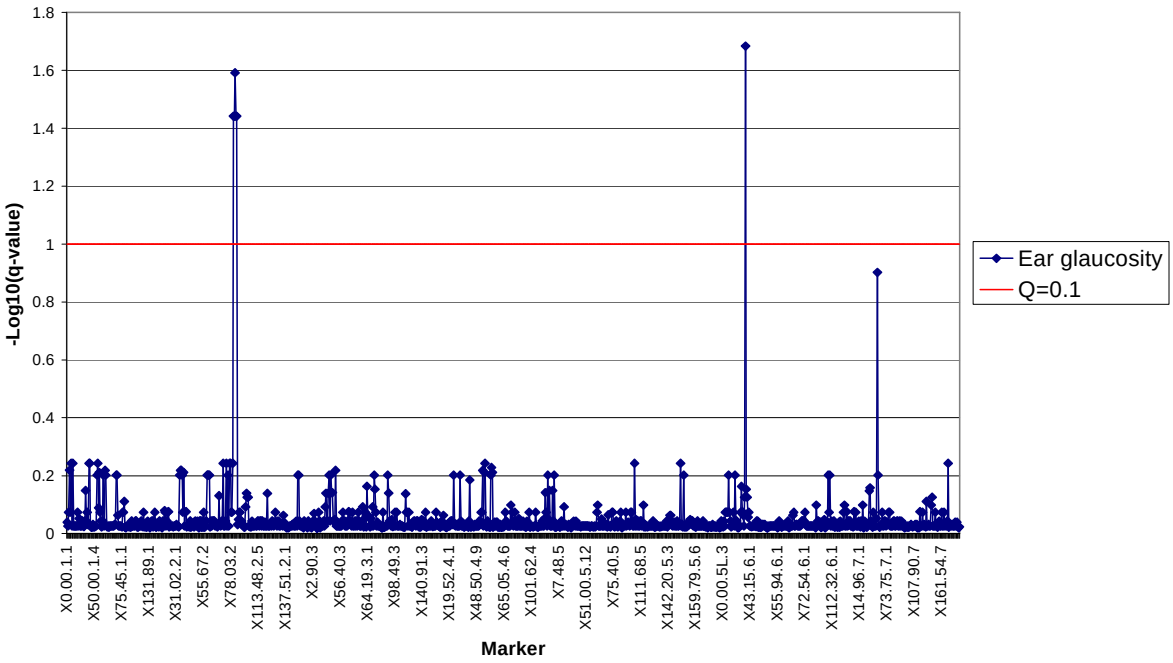


Figure 12-110. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

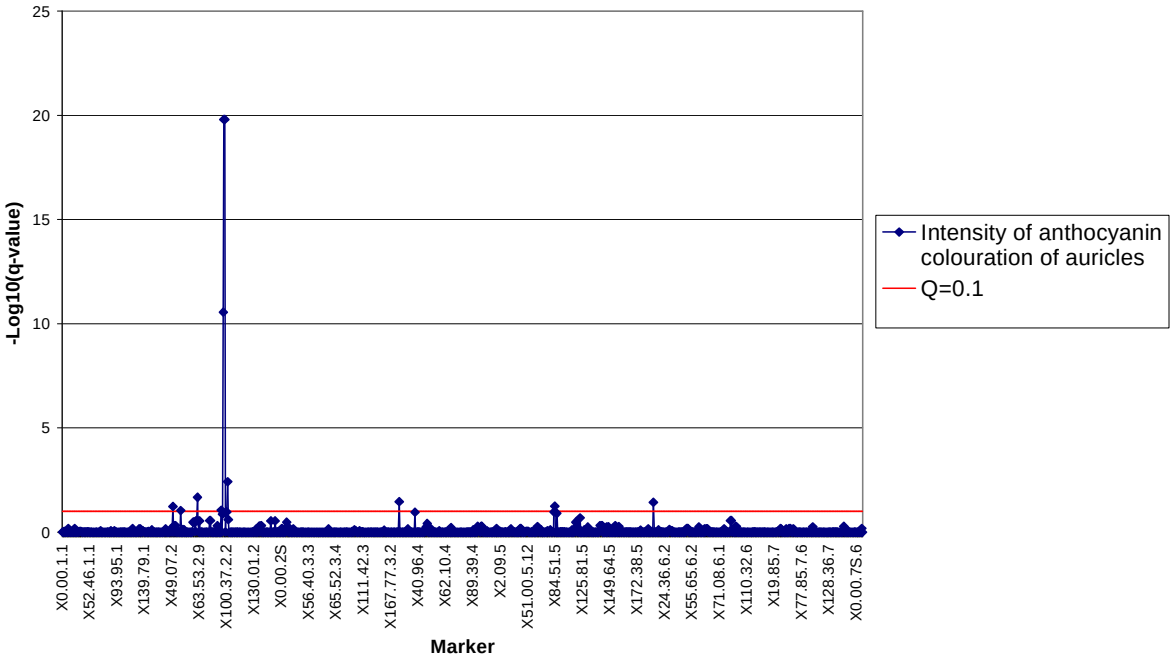


Figure 12-111. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

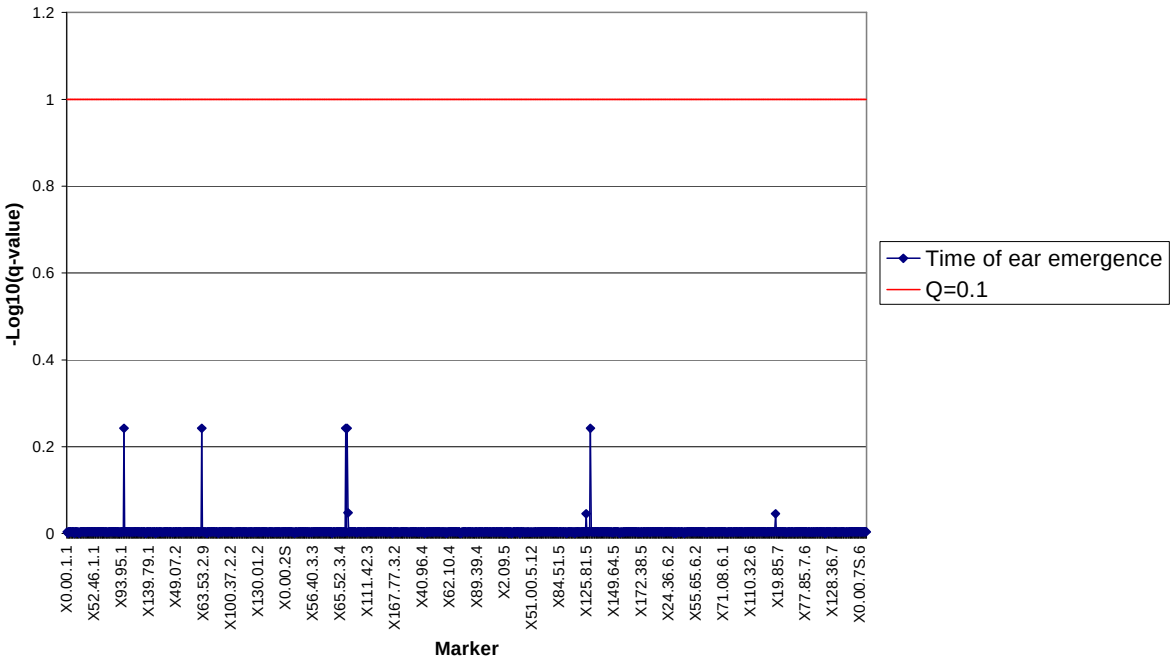


Figure 12-112. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

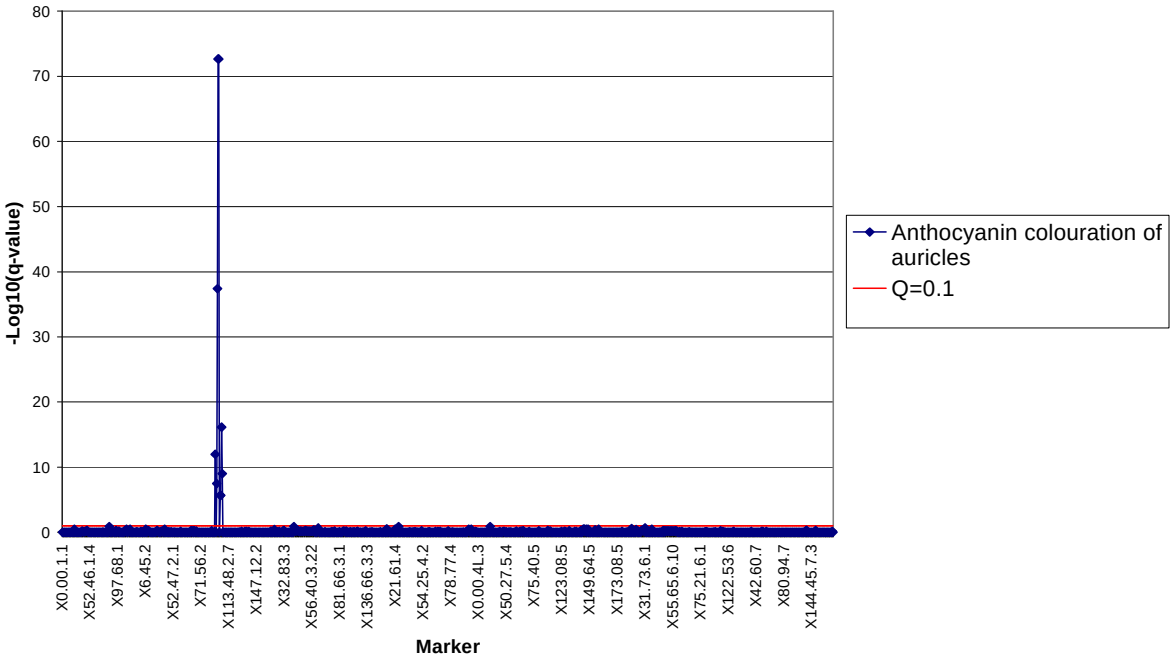


Figure 12-113. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

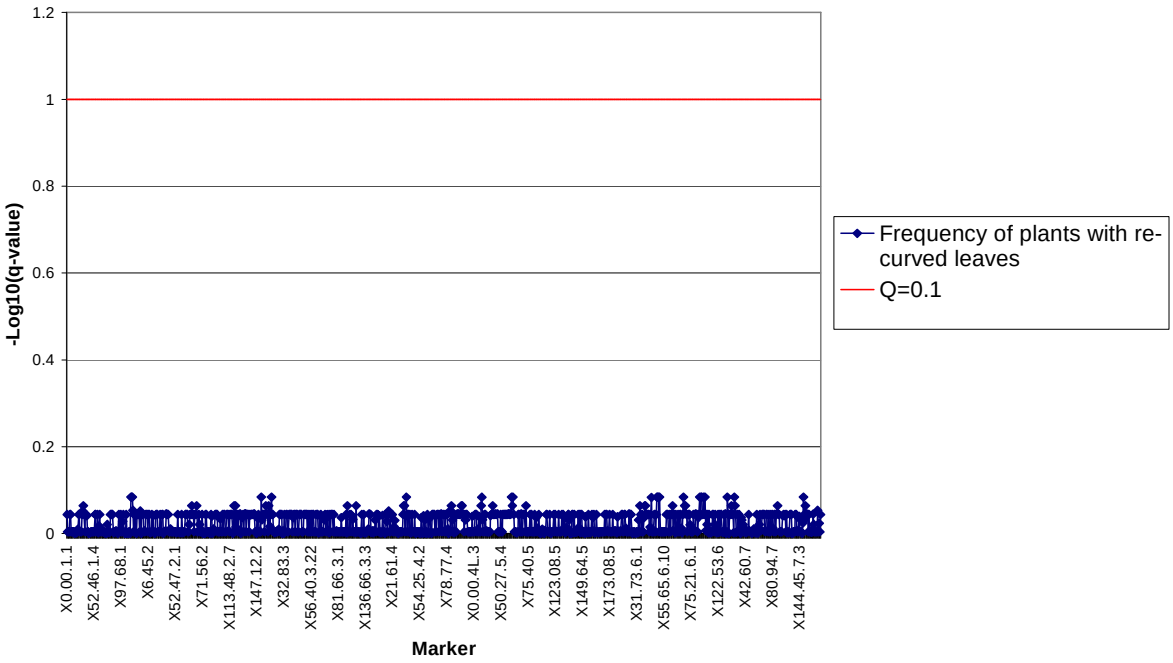


Figure 12-114. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

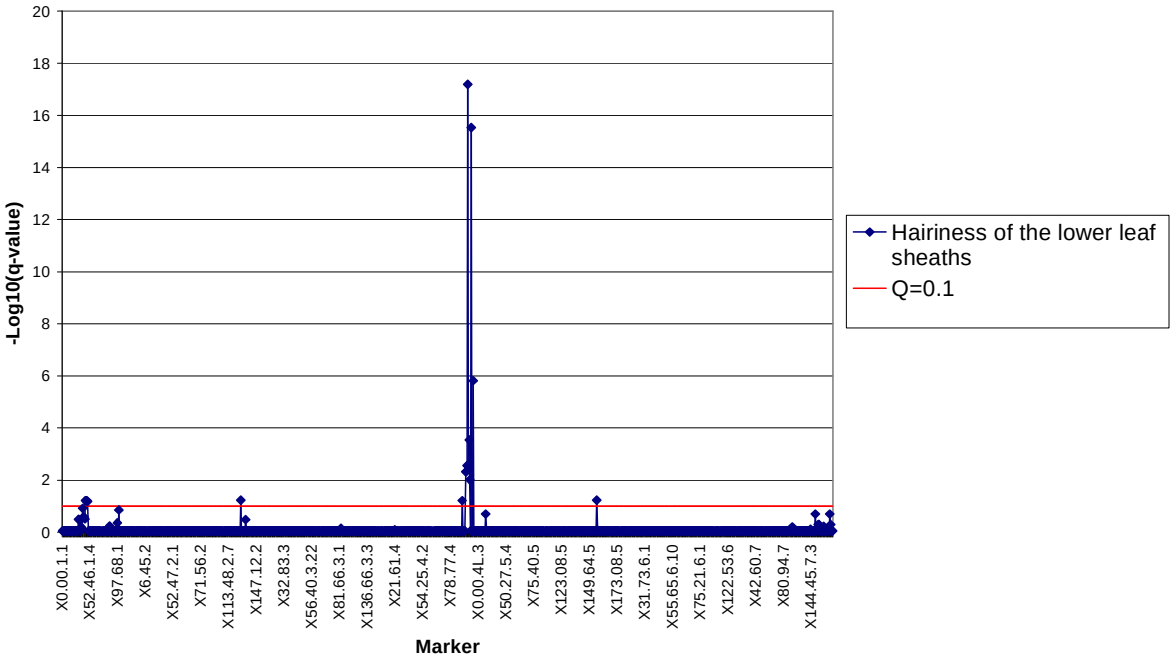


Figure 12-115. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

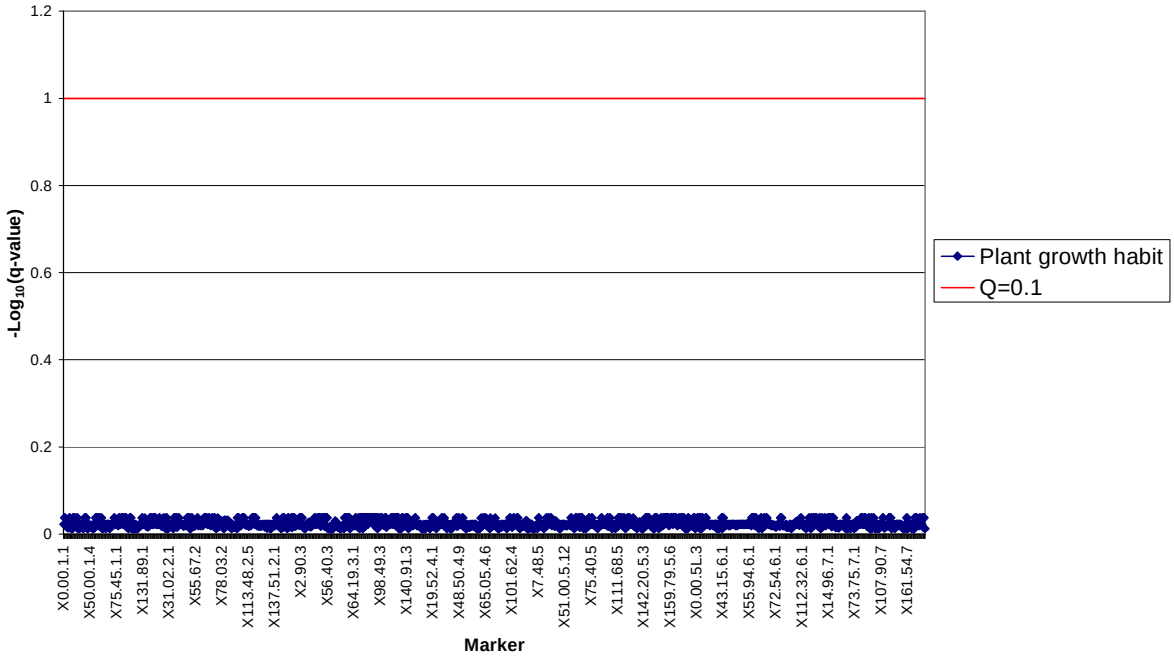


Figure 12-116. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

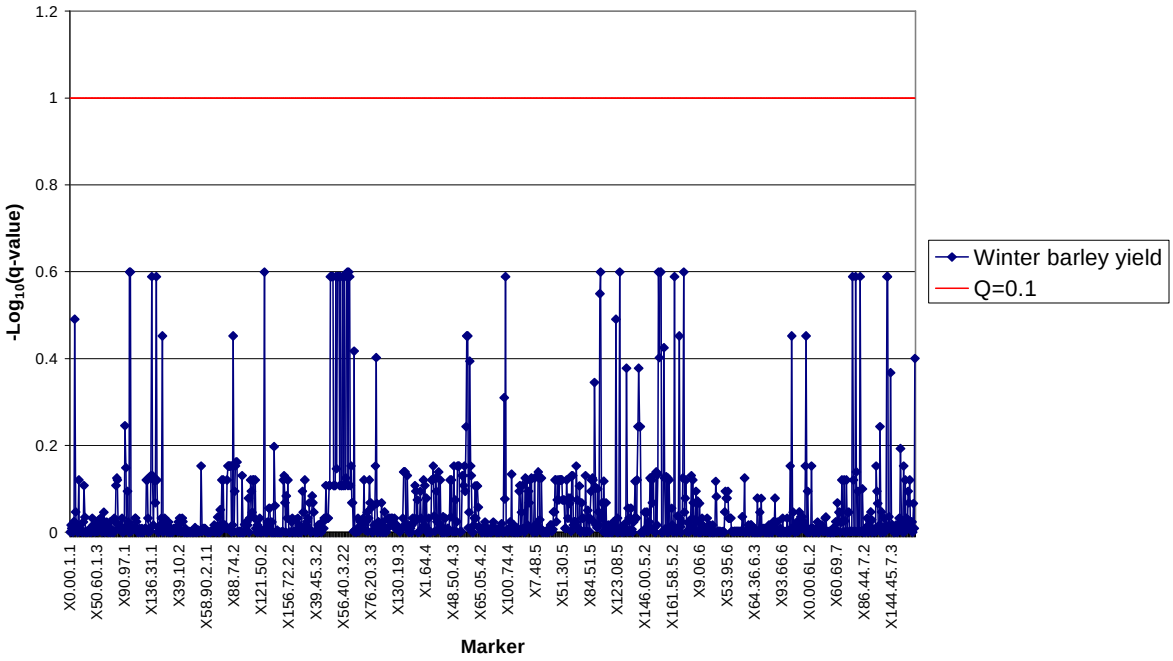


Figure 12-117. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

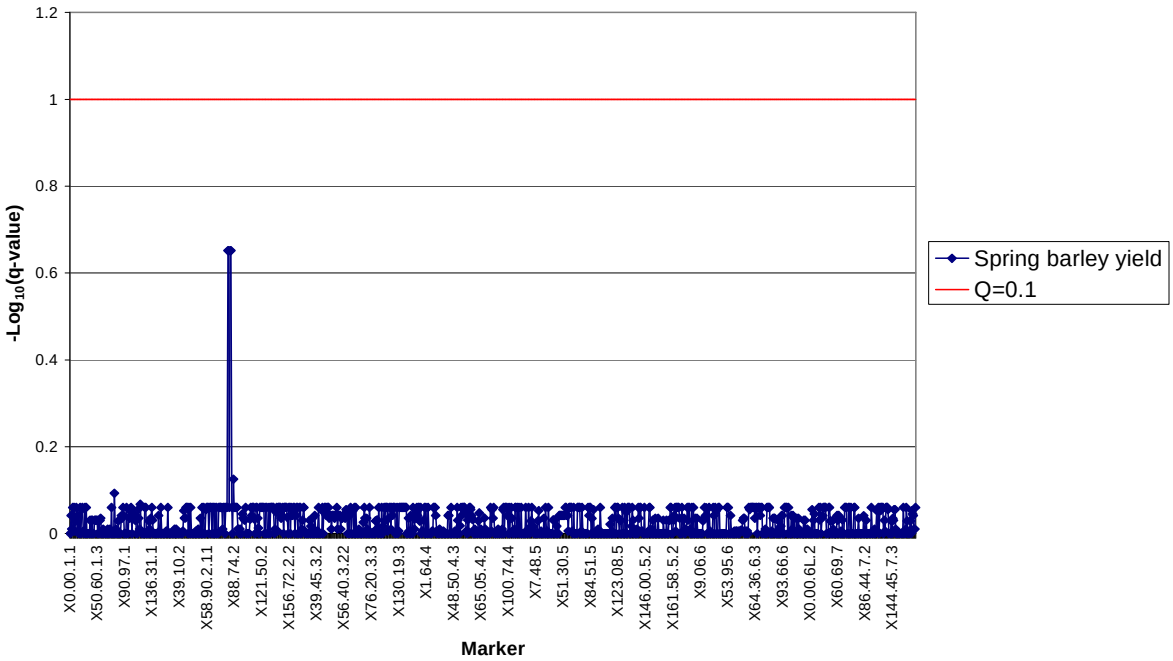


Figure 12-118. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

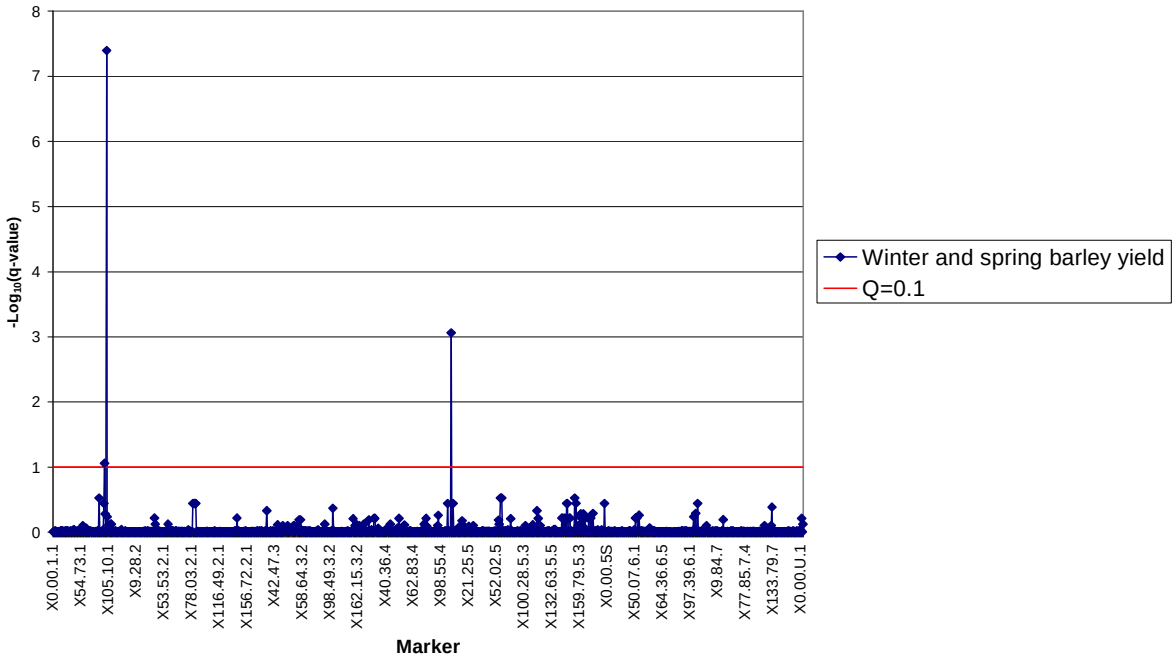


Figure 12-119. Association by PCA + GC. Markers arranged in genetic map order. Trait as indicated in legend

12.10 Appendix 10. Association plots for barley phenotypes. Association estimated with the effects of structure corrected using the mixed model.

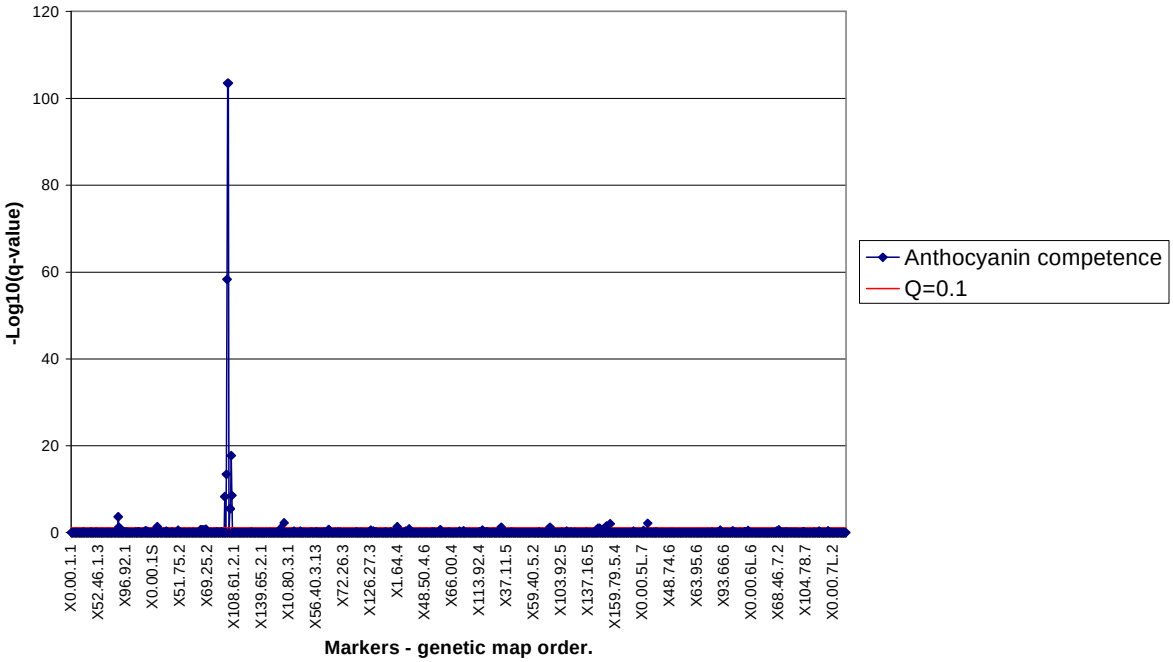


Figure 12-120. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

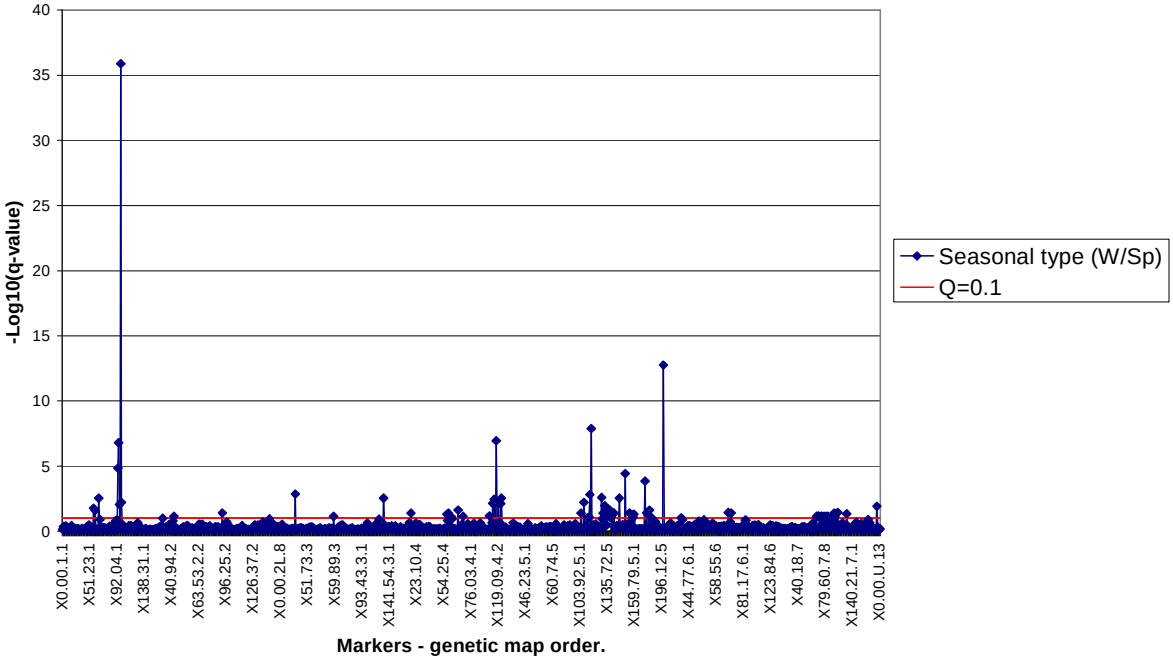


Figure 12-121. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

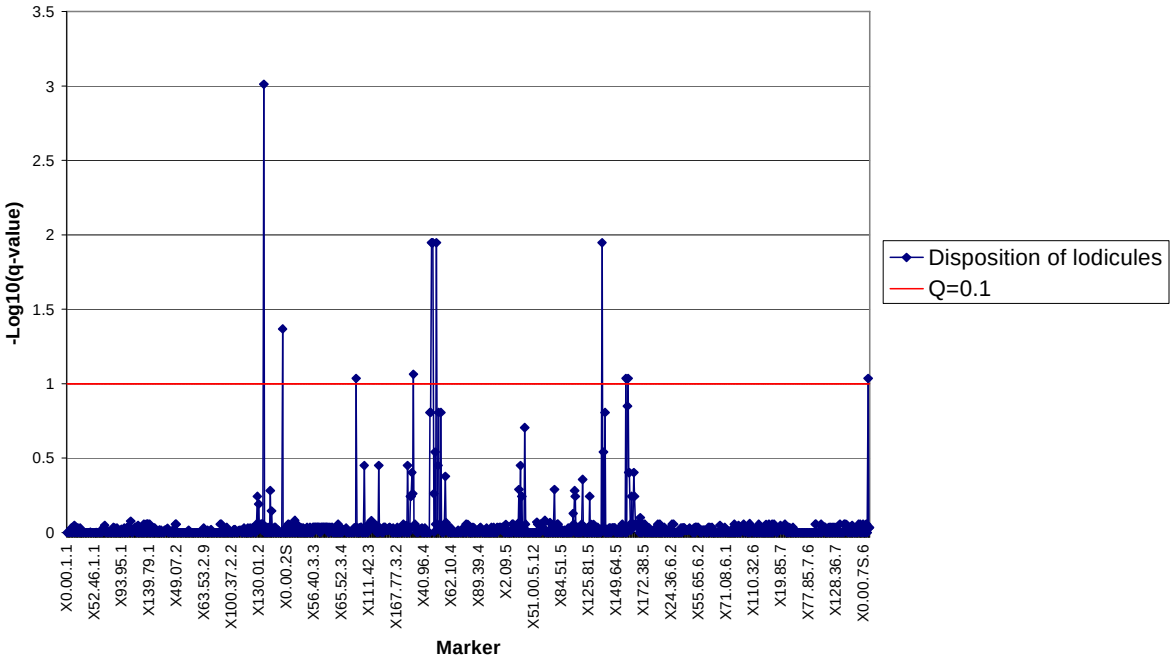


Figure 12-122. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

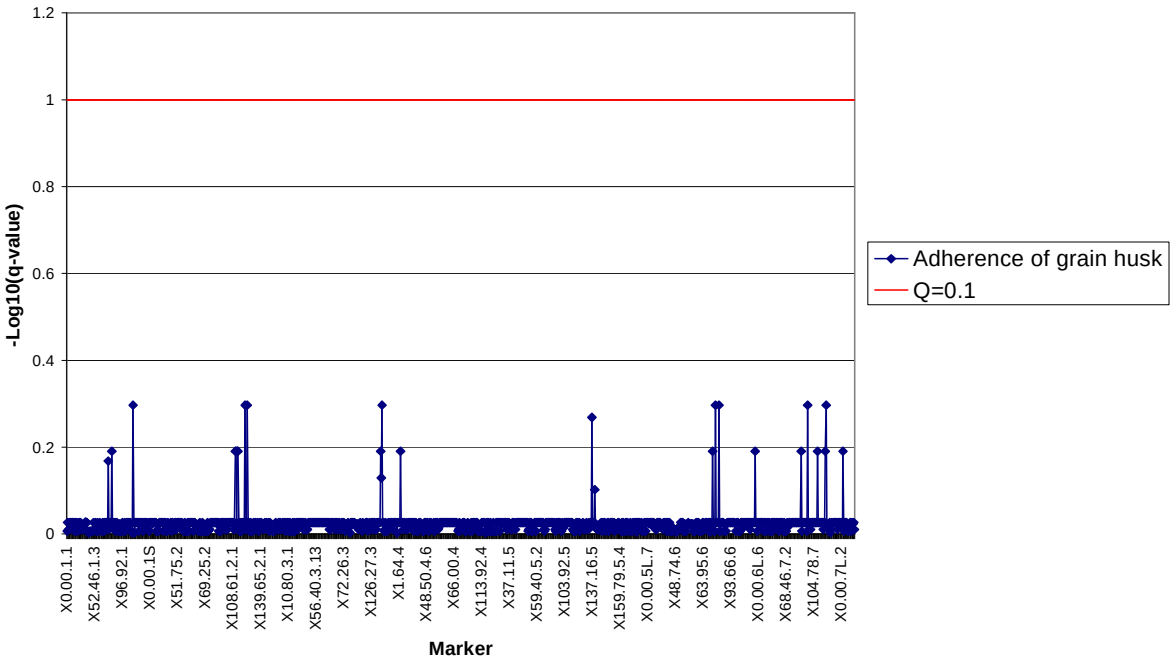


Figure 12-123. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

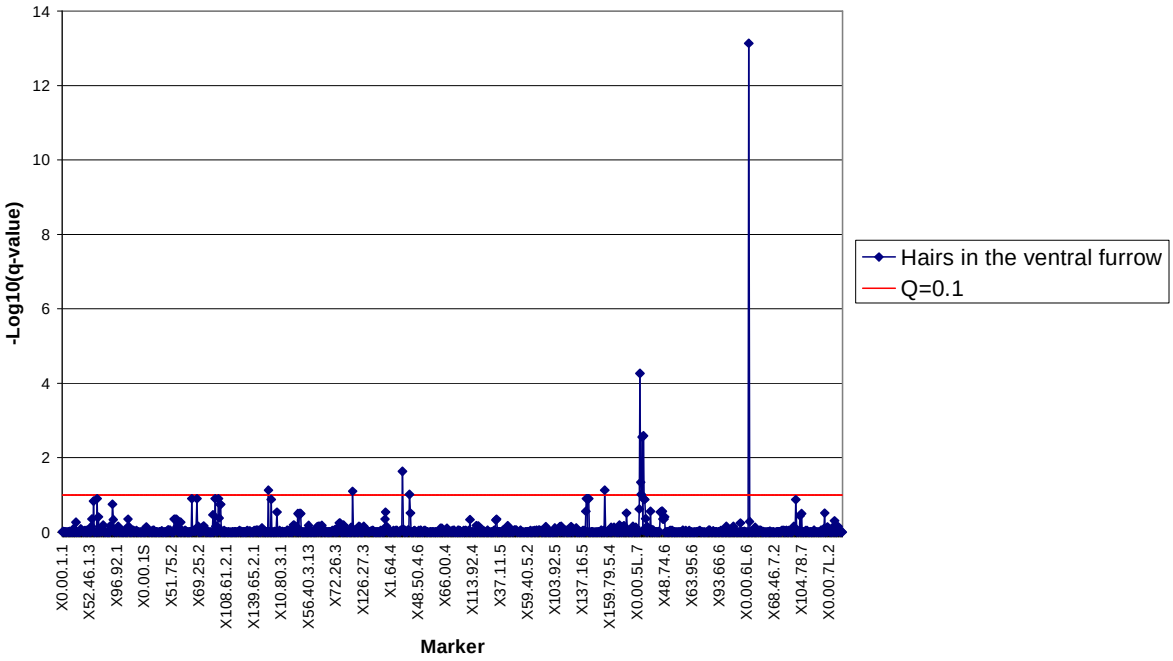


Figure 12-124. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

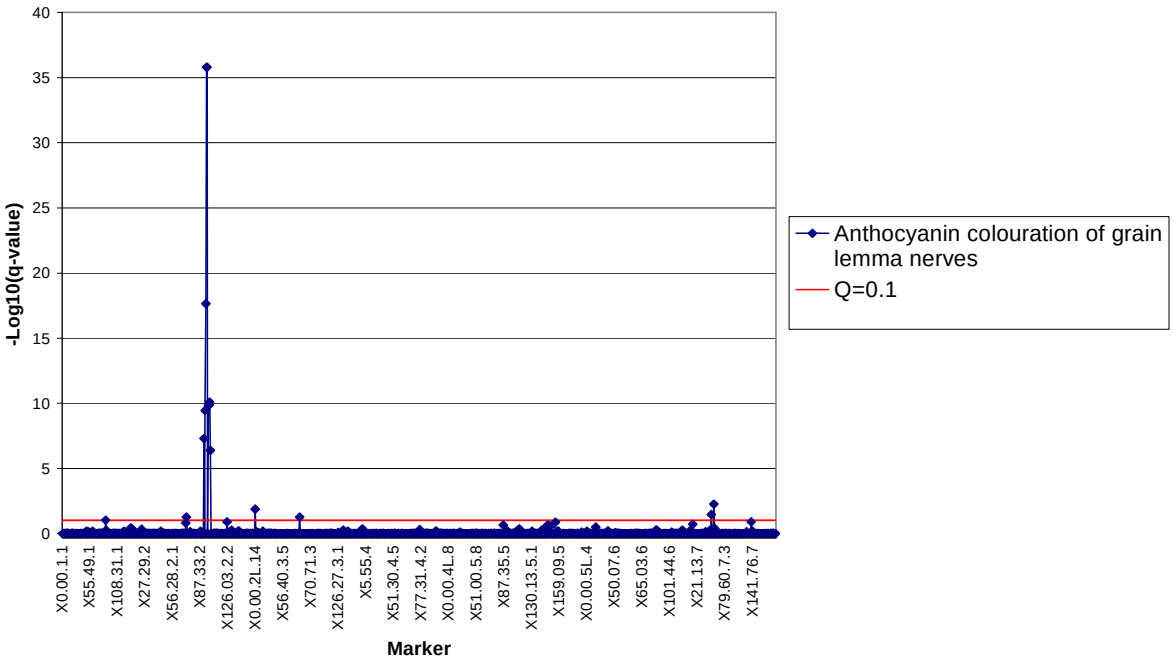


Figure 12-125. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

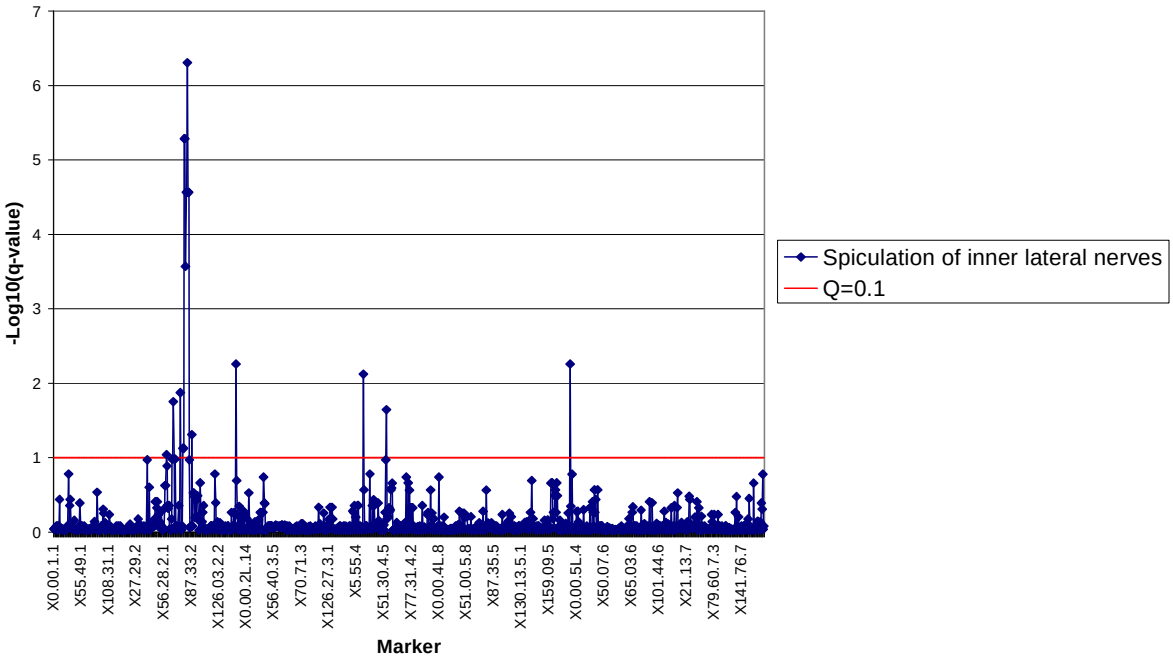


Figure 12-126. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

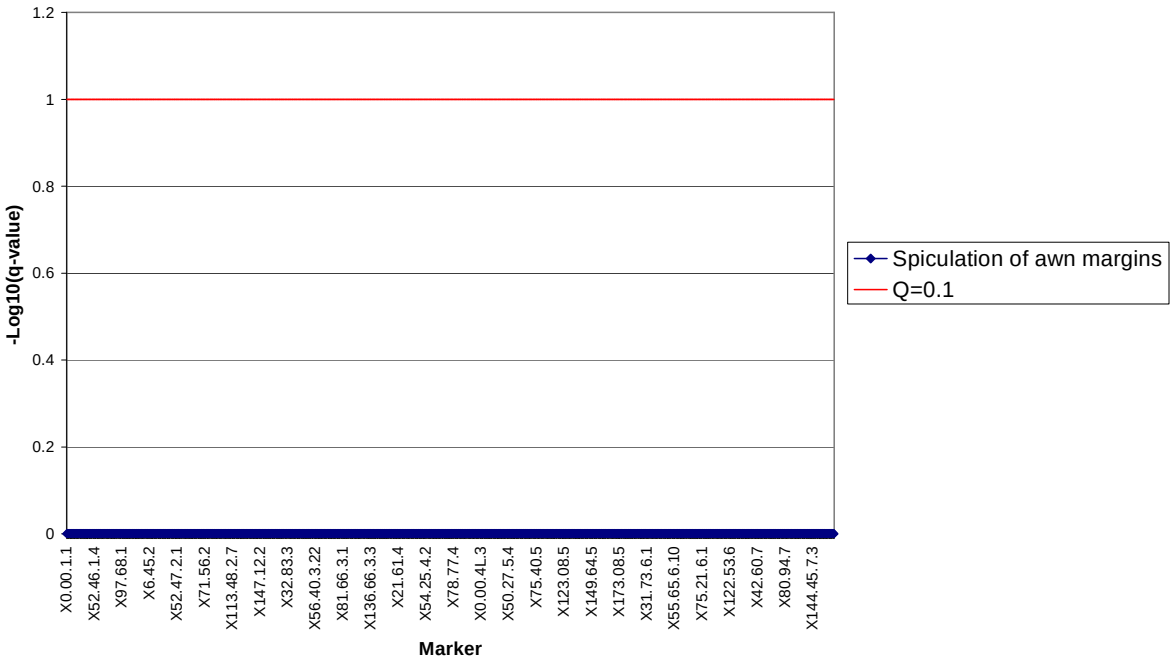


Figure 12-127. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

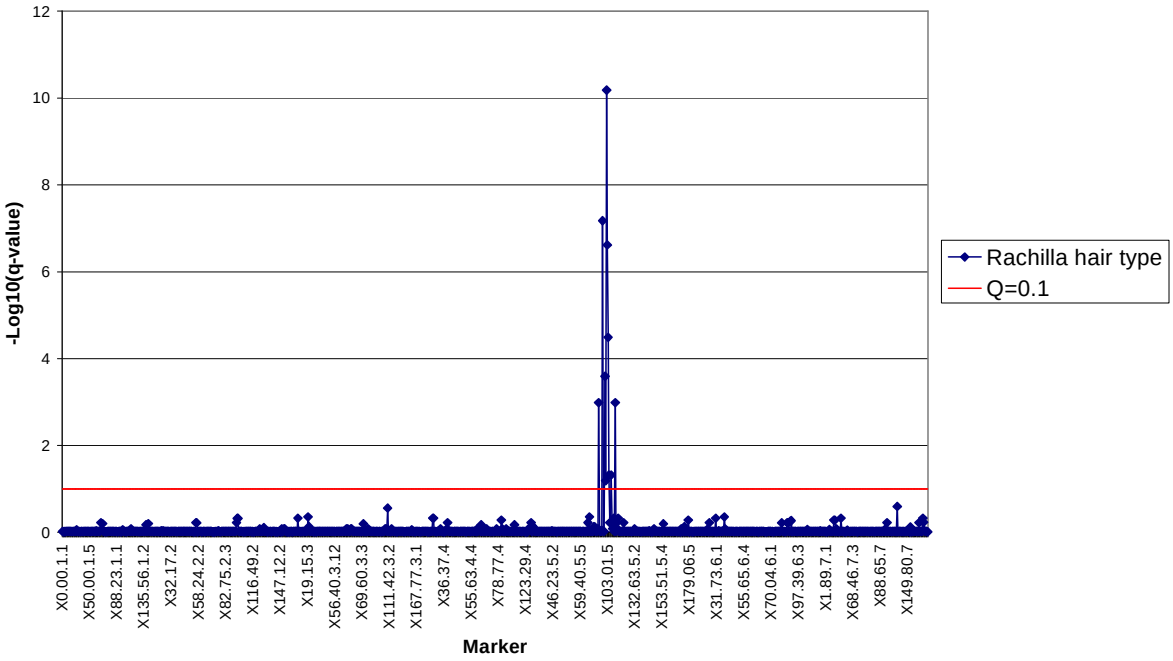


Figure 12-128. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

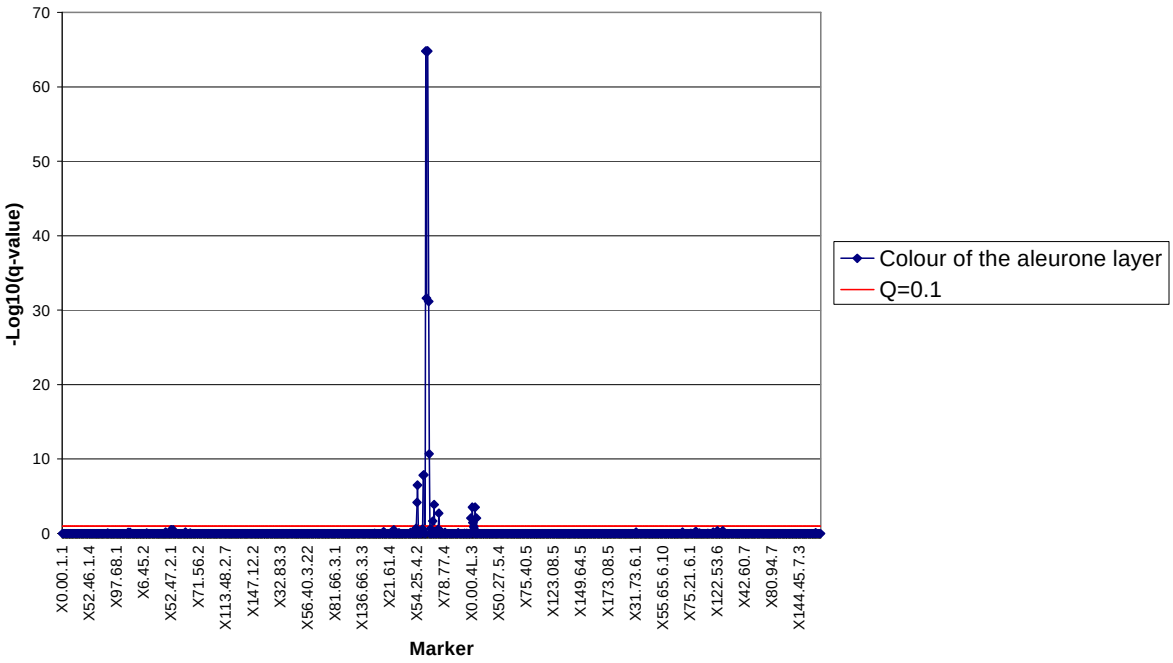


Figure 12-129. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

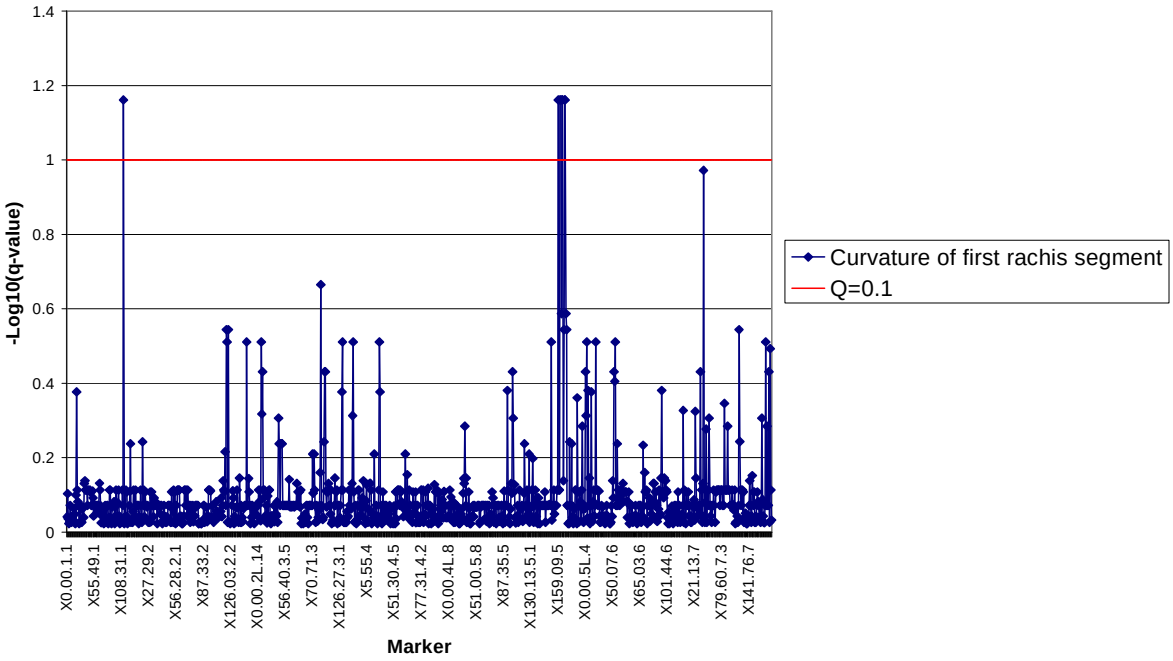


Figure 12-130. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

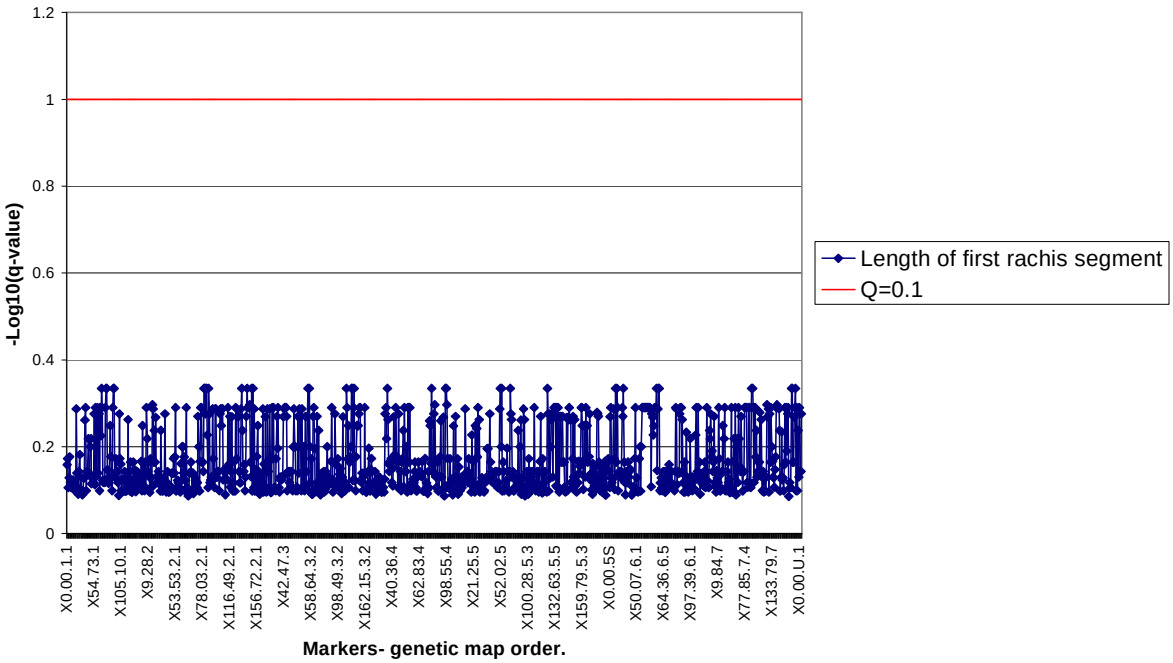


Figure 12-131. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

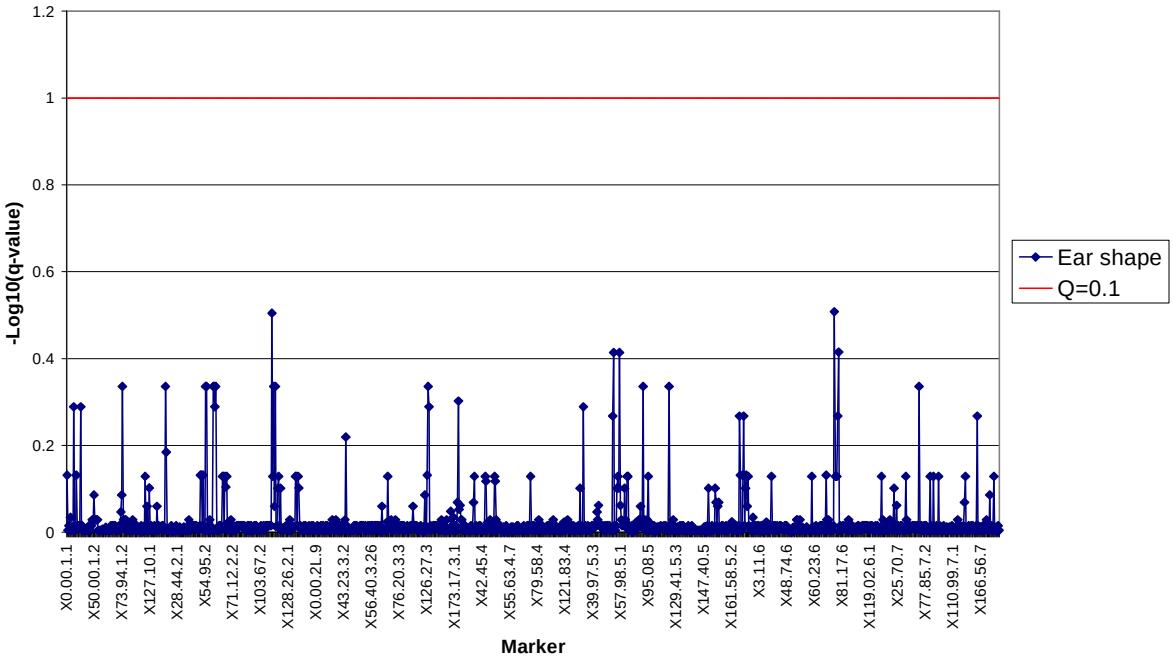


Figure 12-132. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

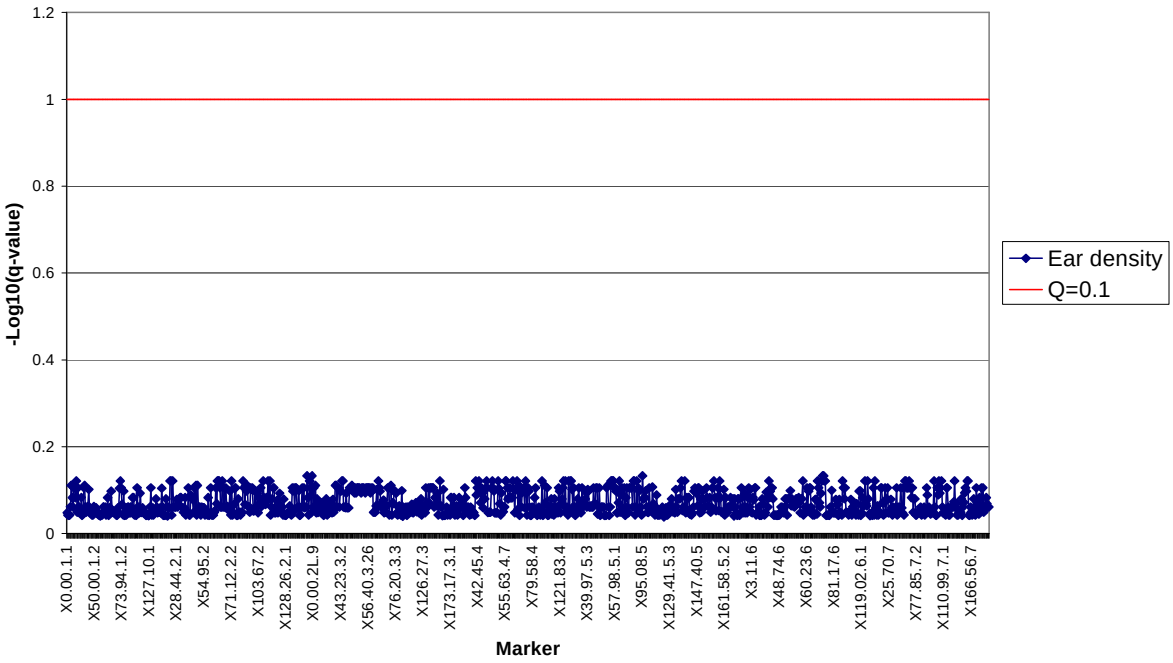


Figure 12-133. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

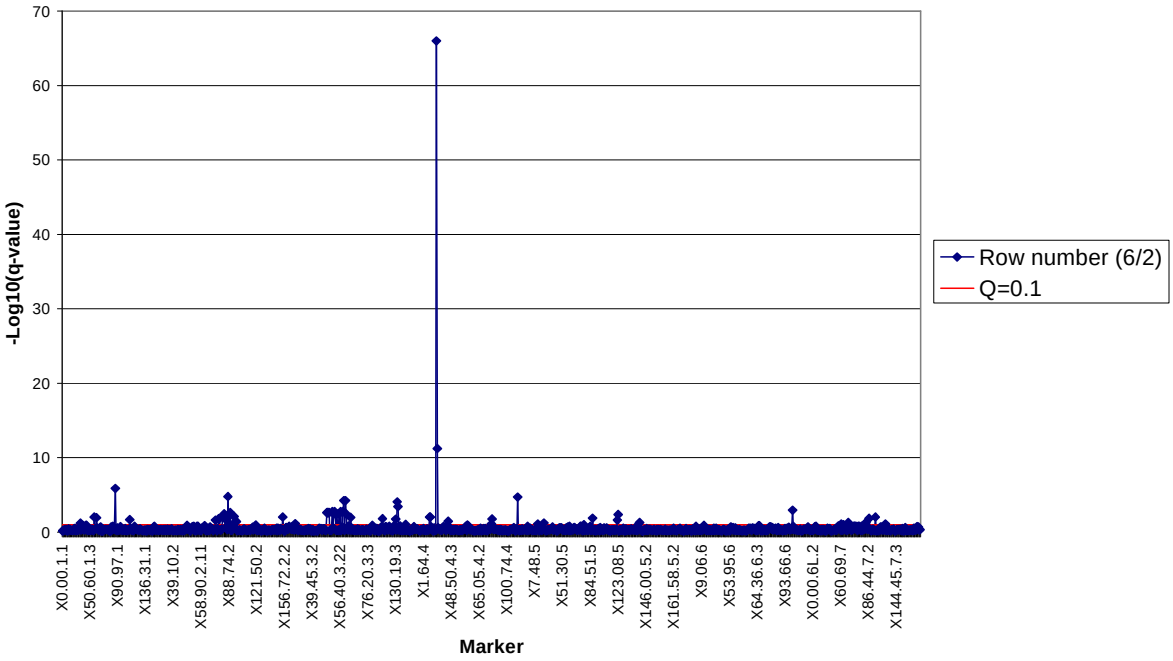


Figure 12-134. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

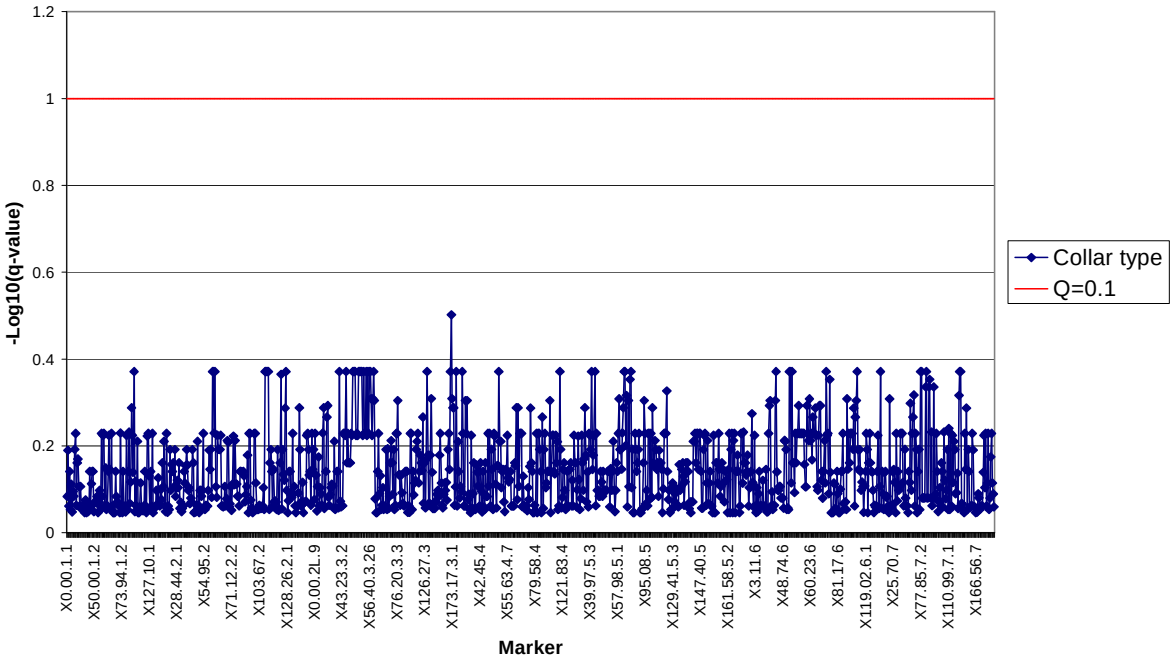


Figure 12-135. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

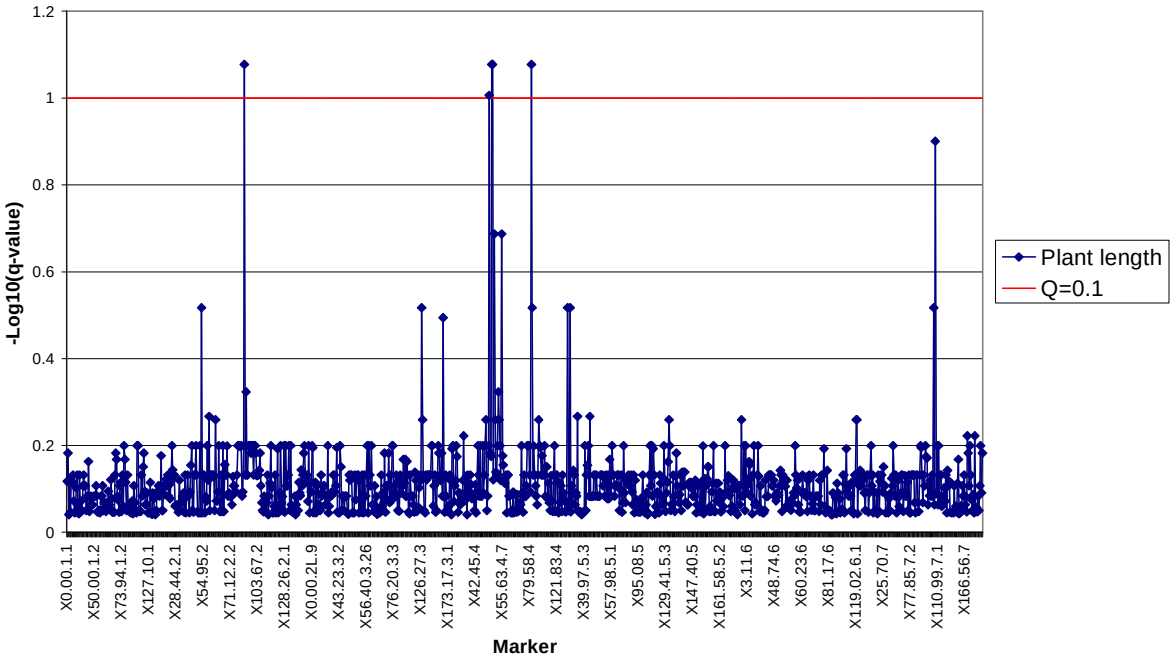


Figure 12-136. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

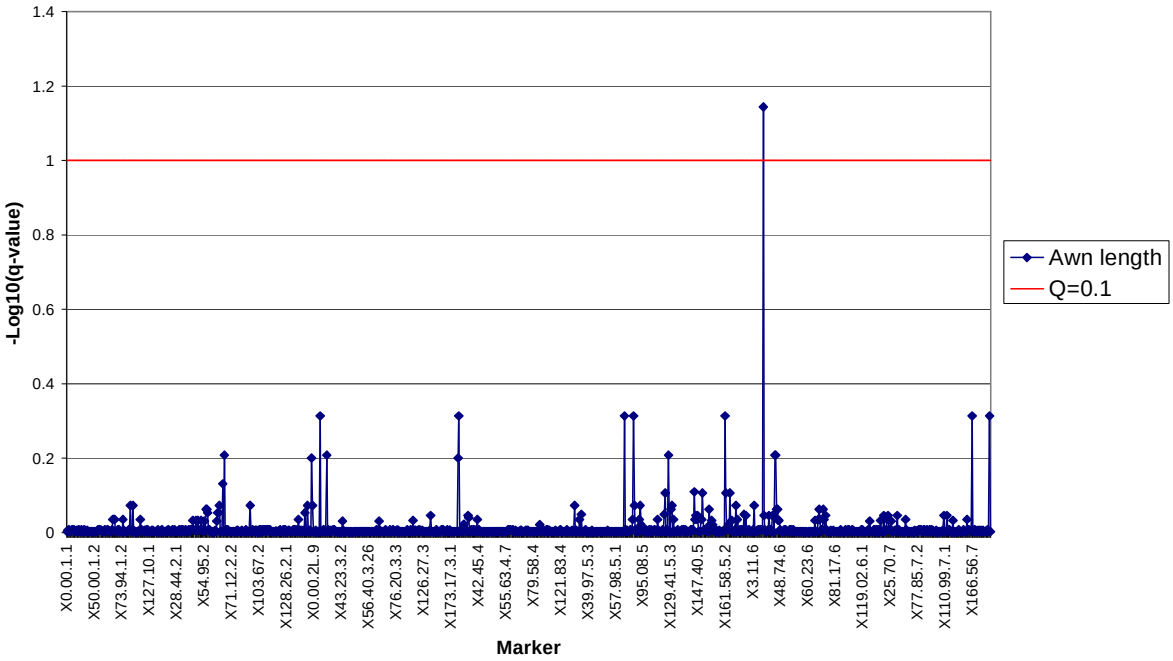


Figure 12-137. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

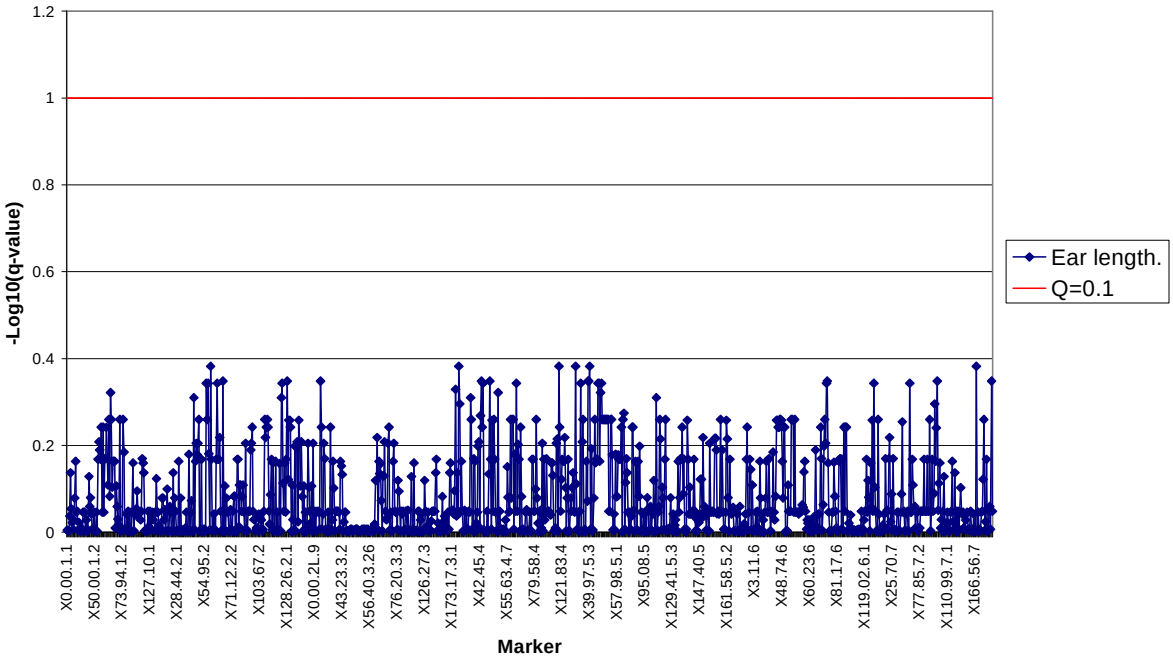


Figure 12-138. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

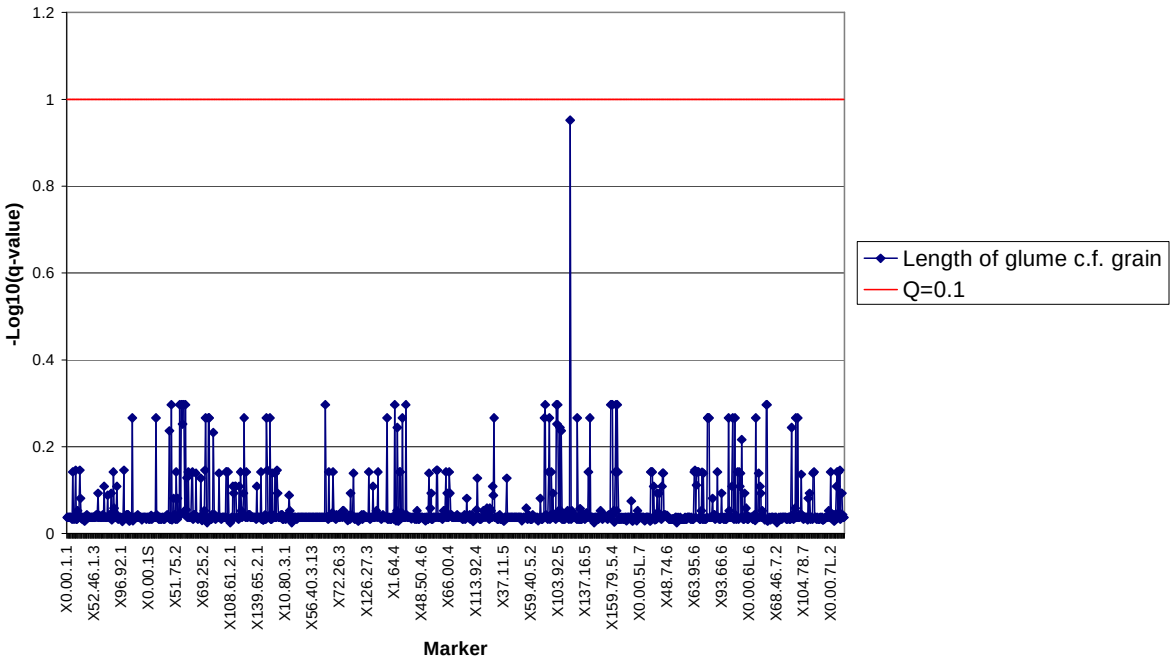


Figure 12-139. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

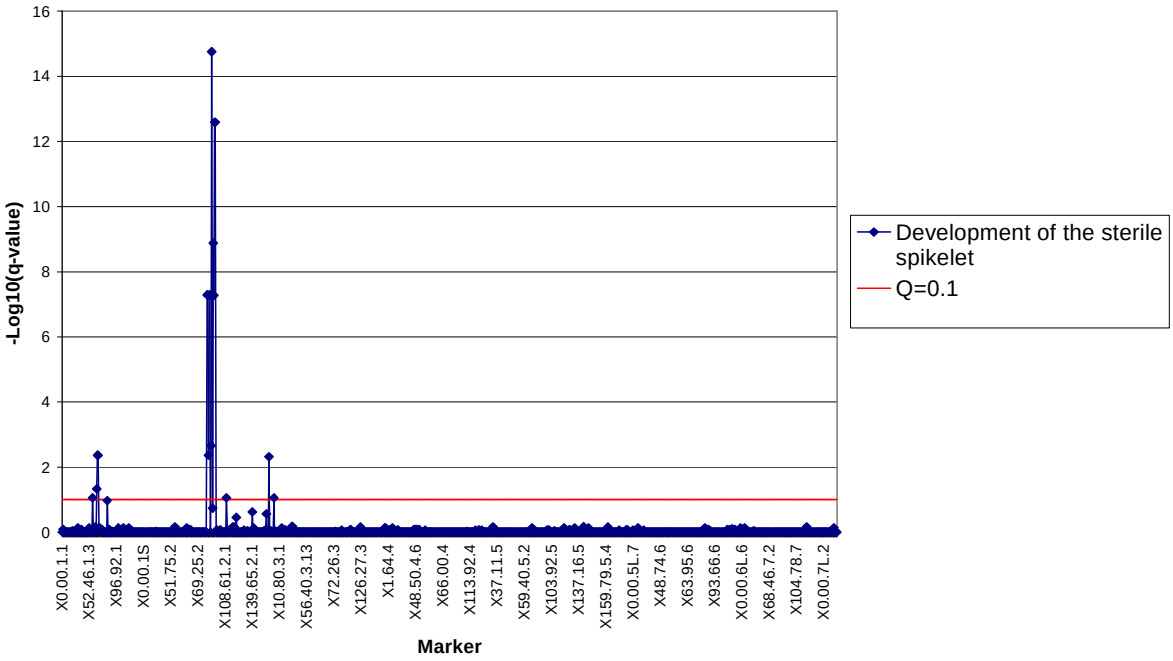


Figure 12-140. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

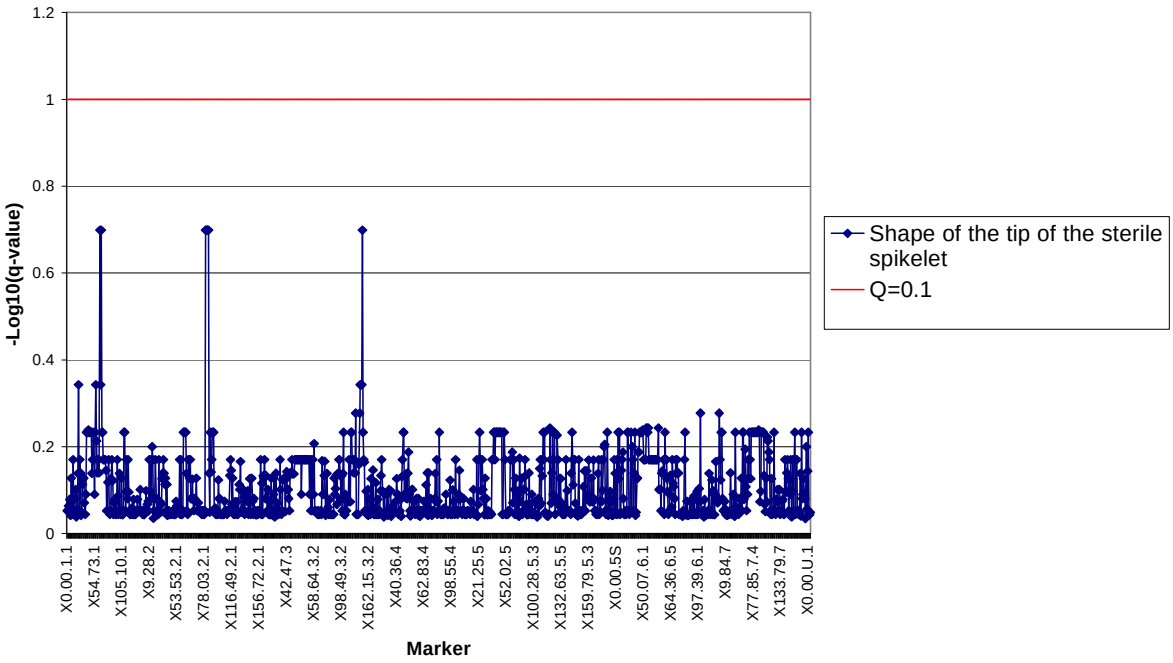


Figure 12-141. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

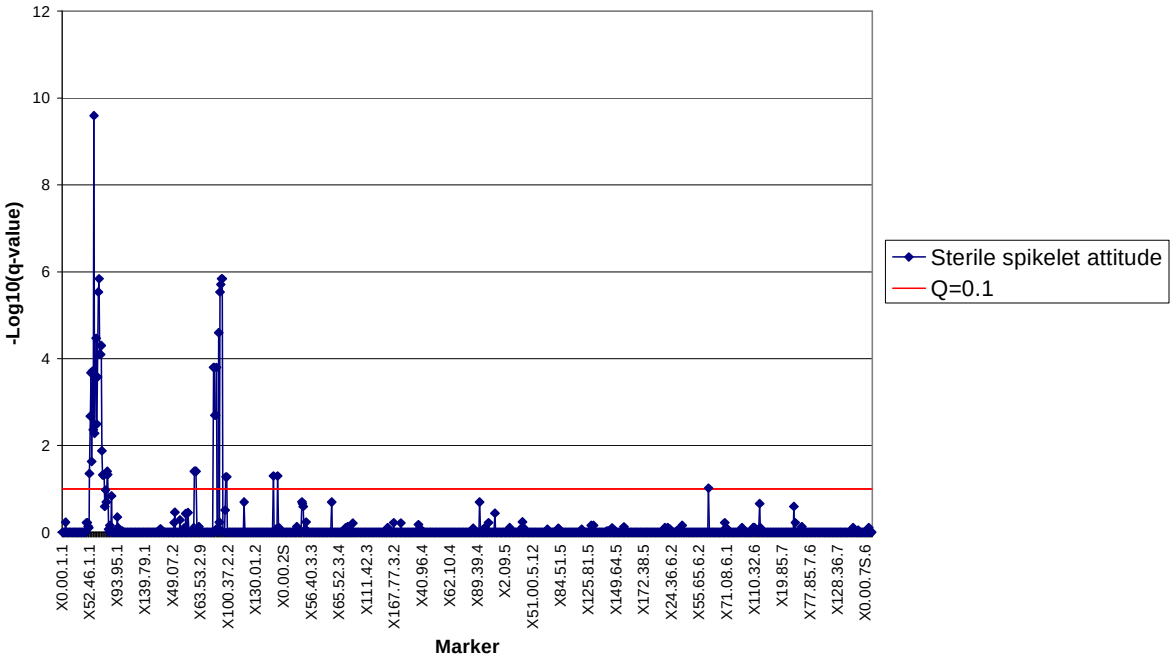


Figure 12-142. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

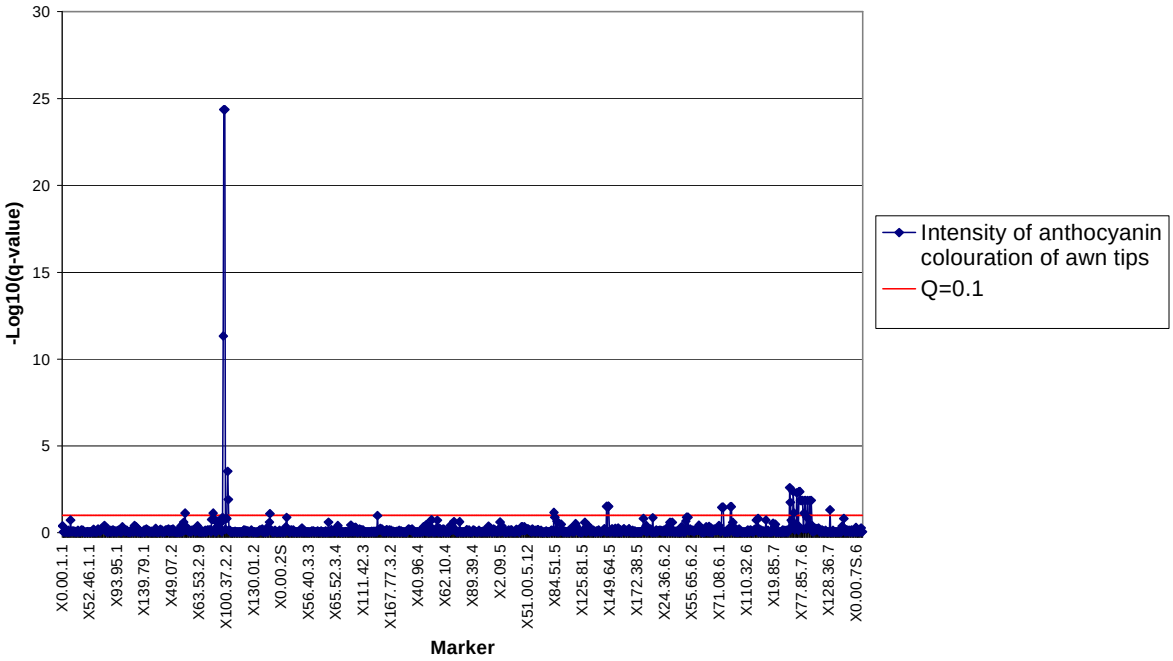


Figure 12-143. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

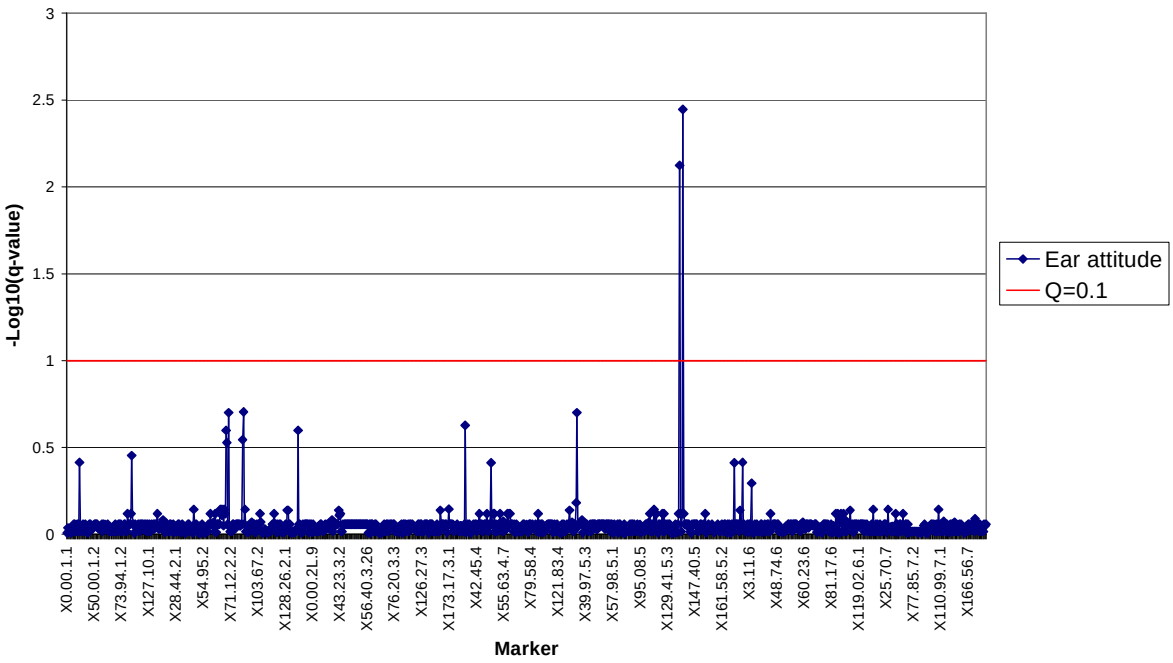


Figure 12-144. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

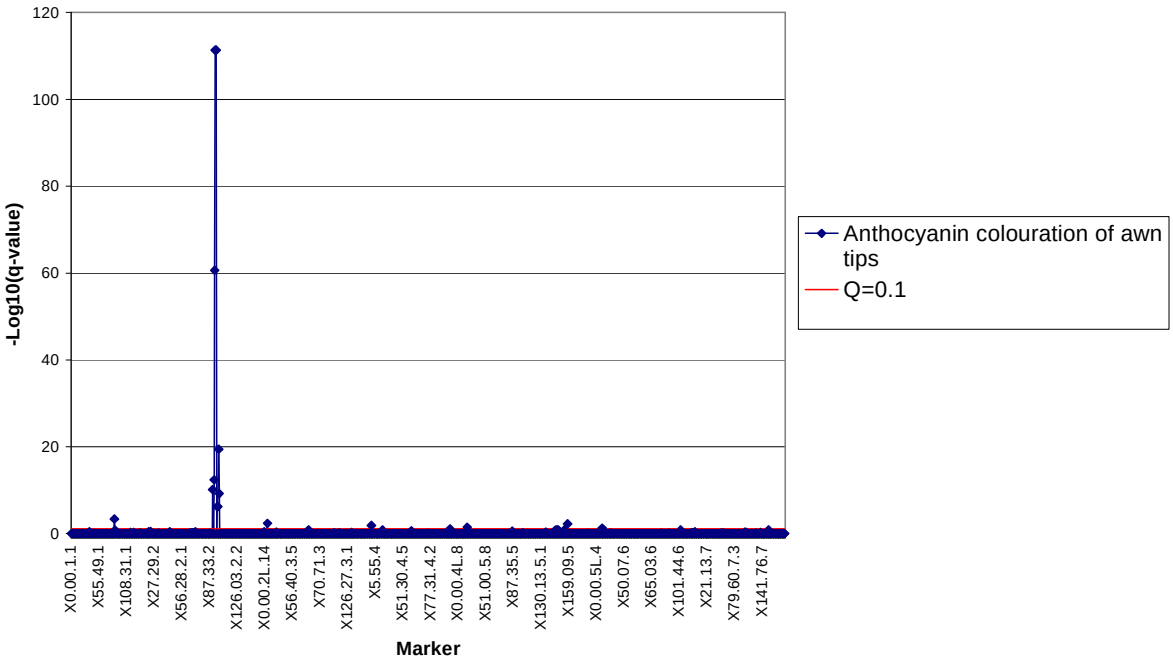


Figure 12-145. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

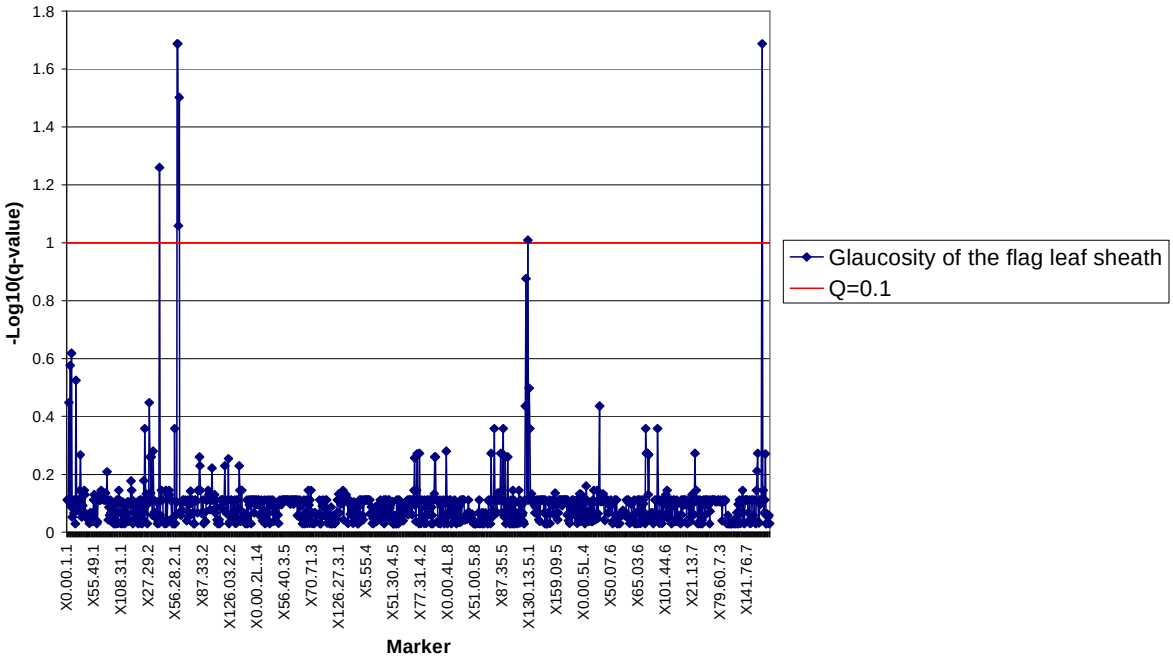


Figure 12-146. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

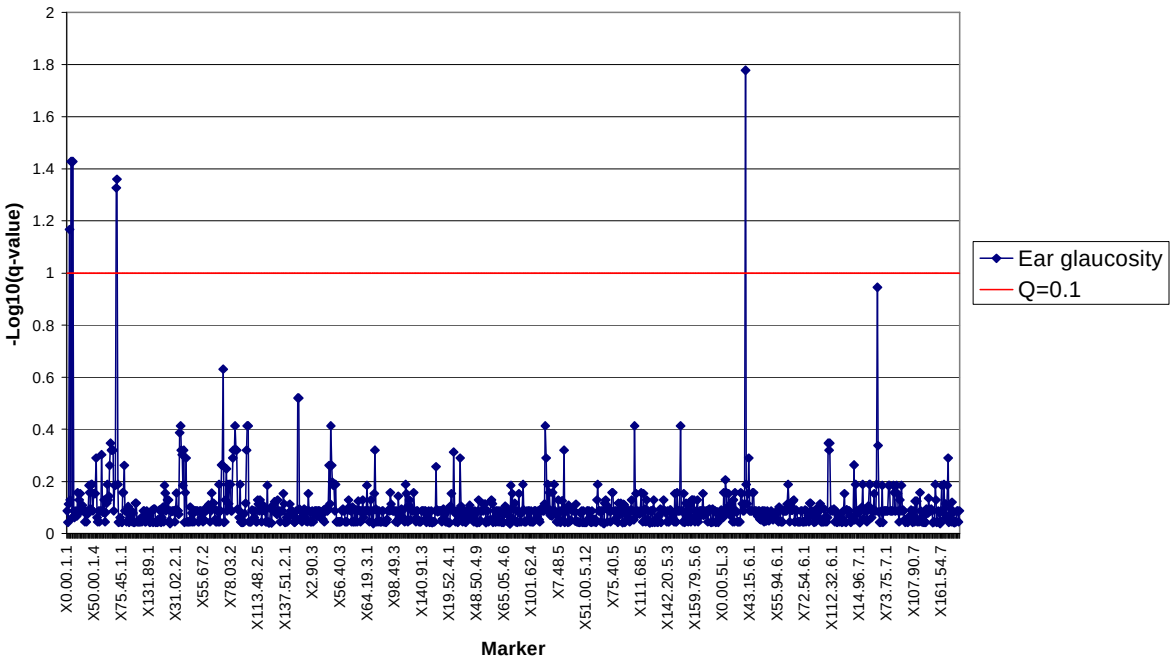


Figure 12-147. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

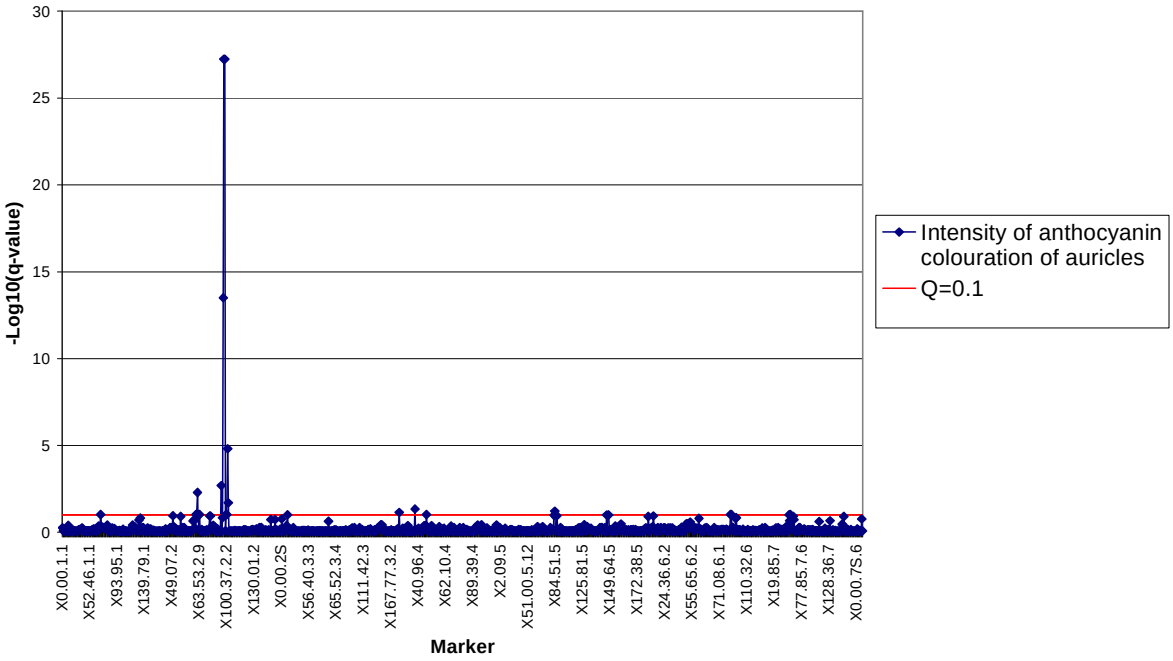


Figure 12-148. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

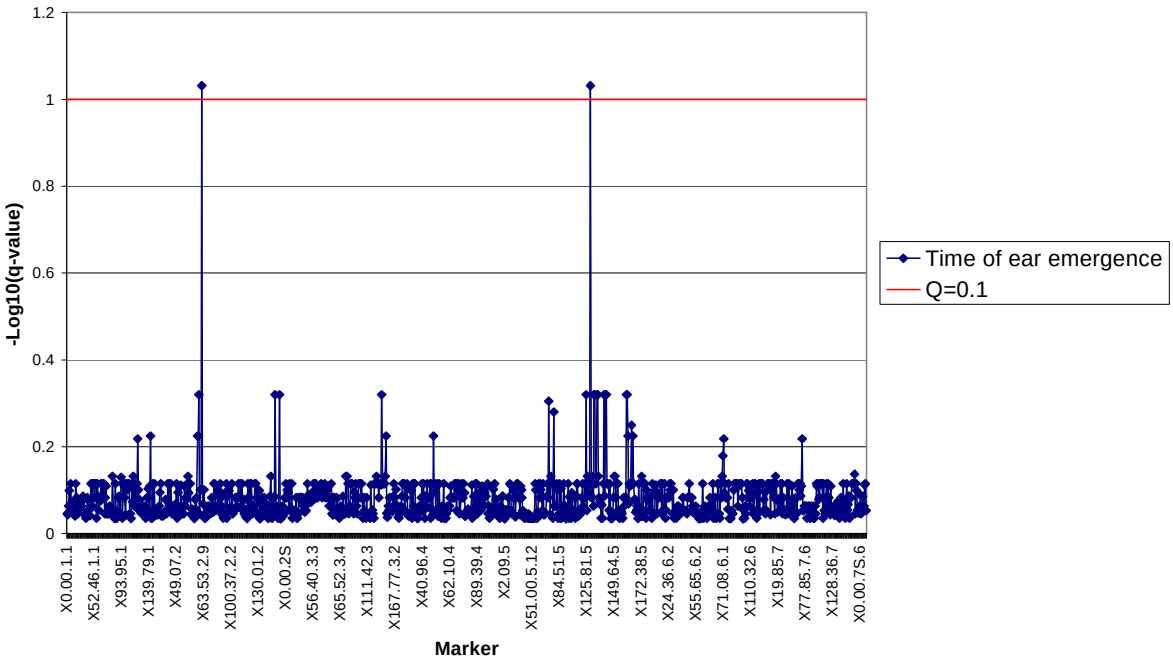


Figure 12-149. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

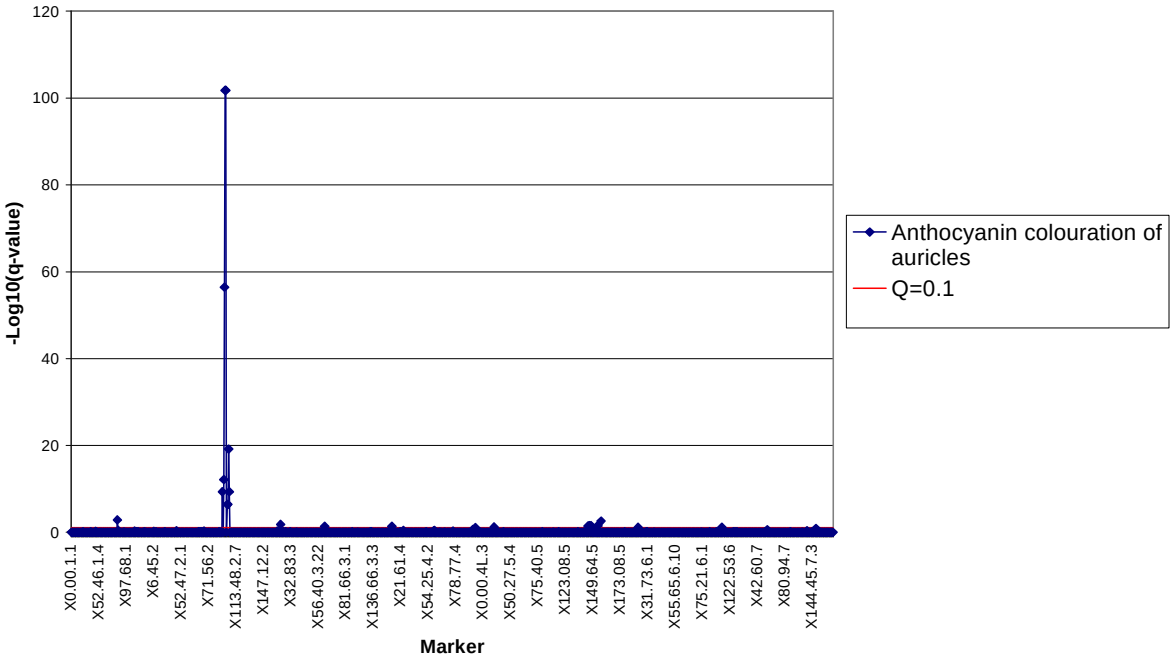


Figure 12-150. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

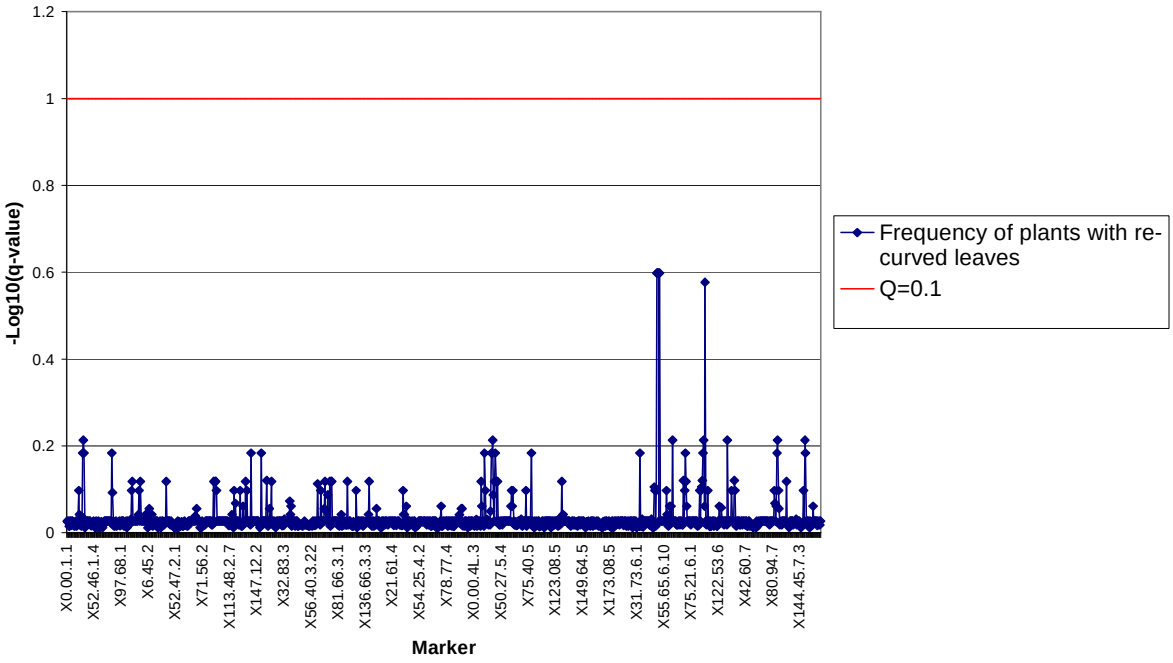


Figure 12-151. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

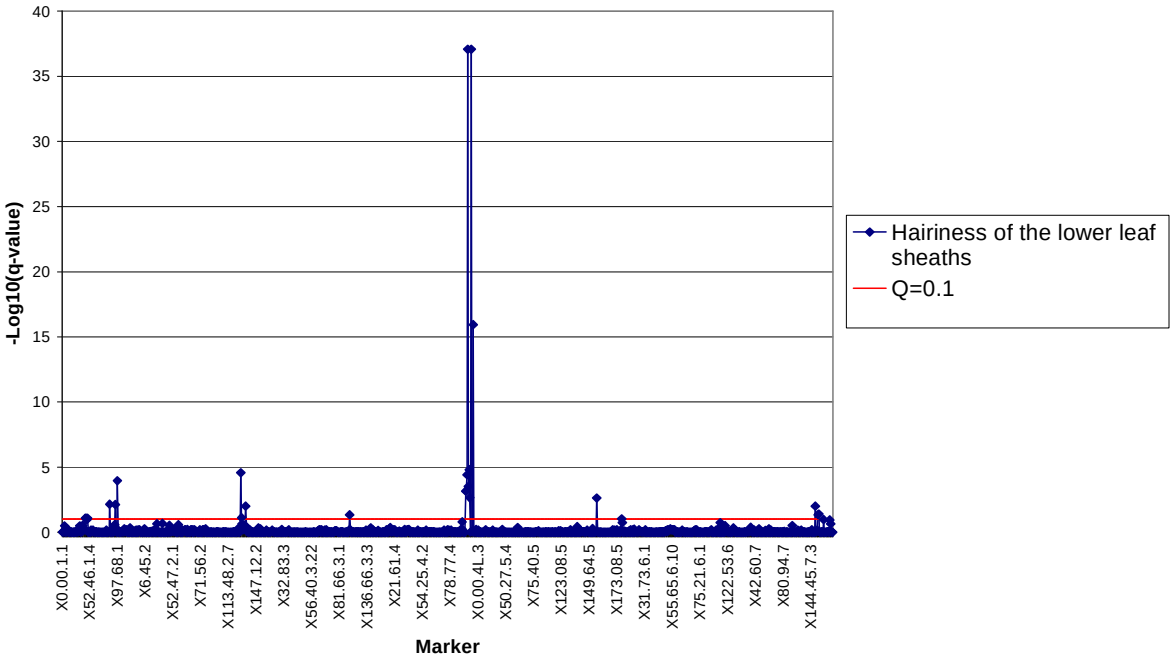


Figure 12-152. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

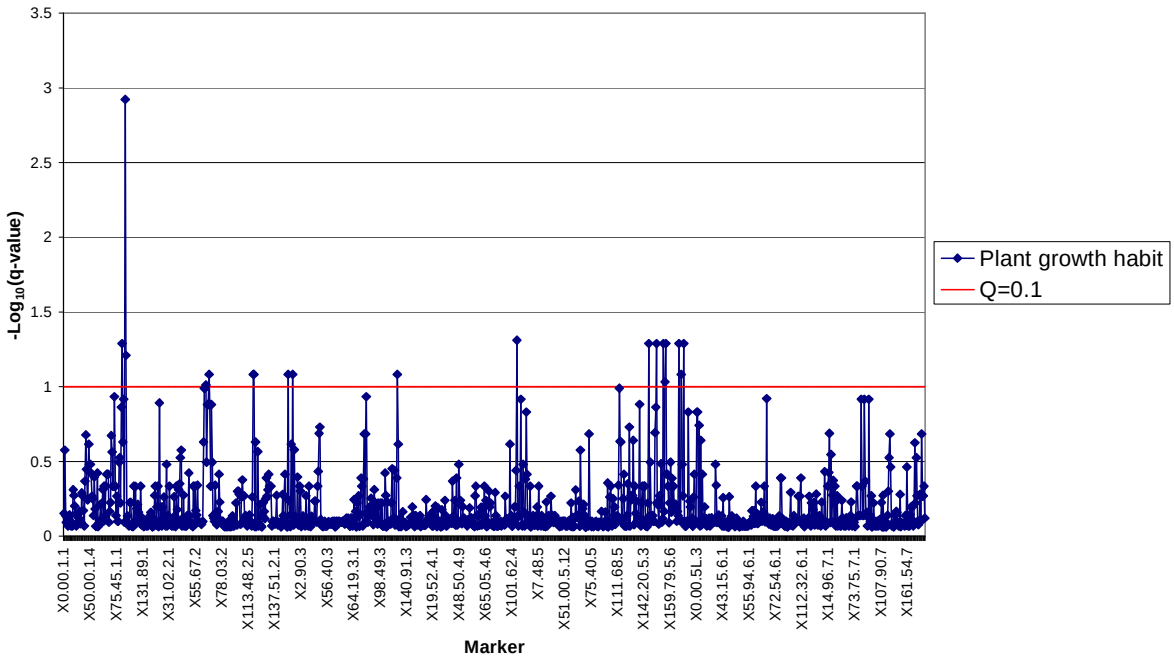


Figure 12-153. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

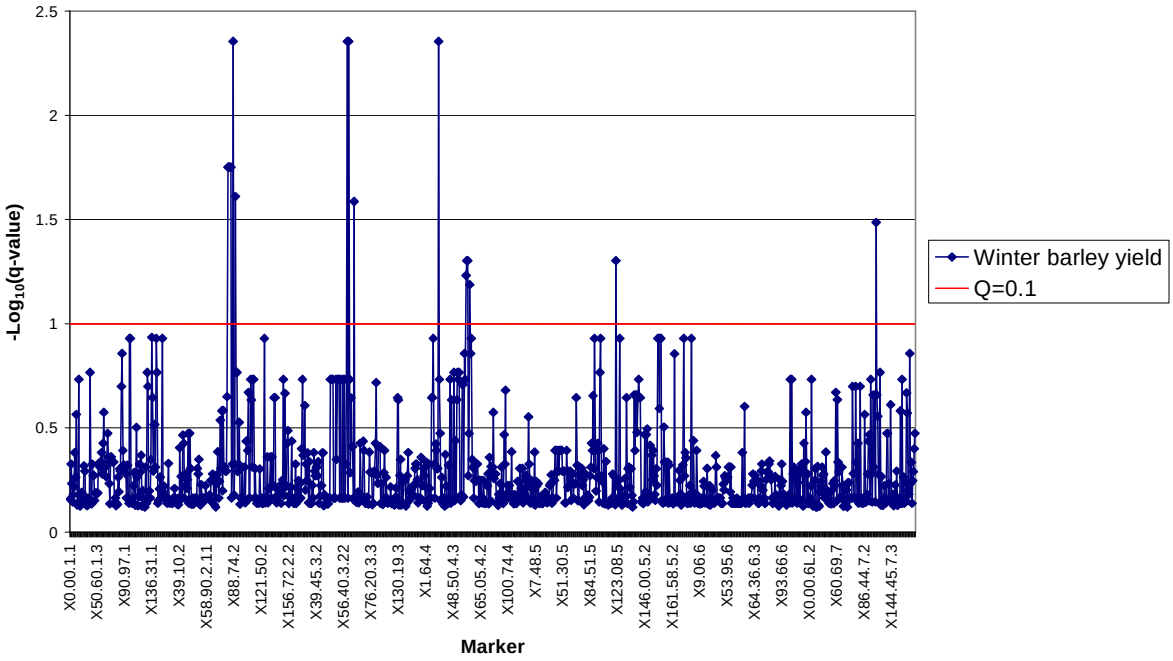


Figure 12-154. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

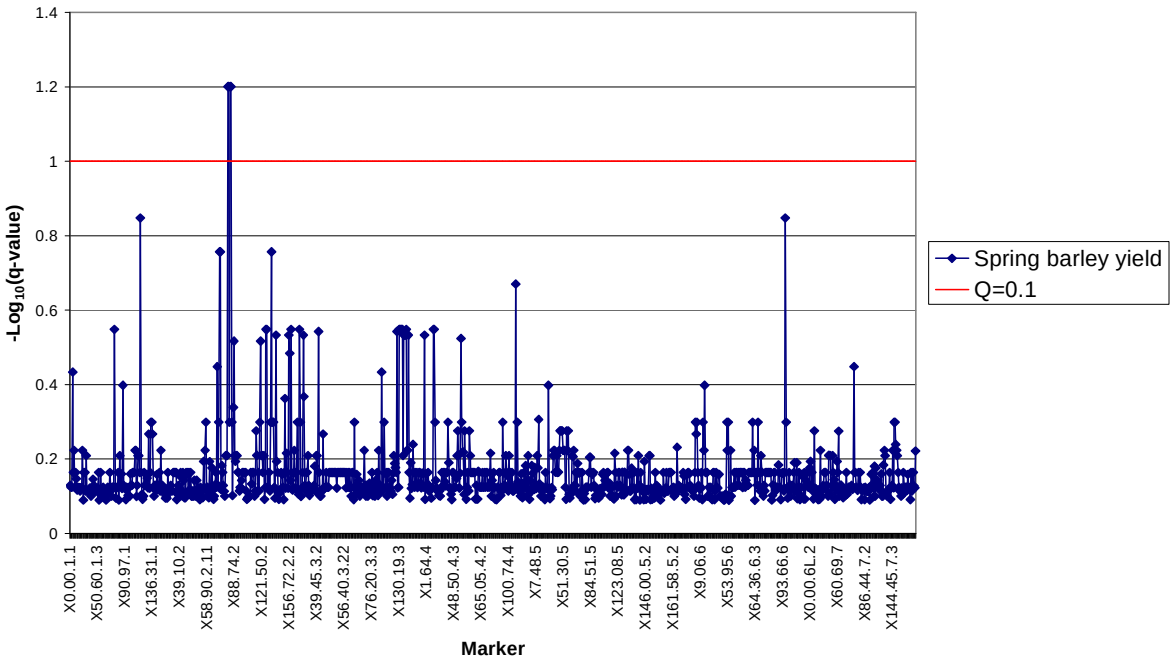


Figure 12-155. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend

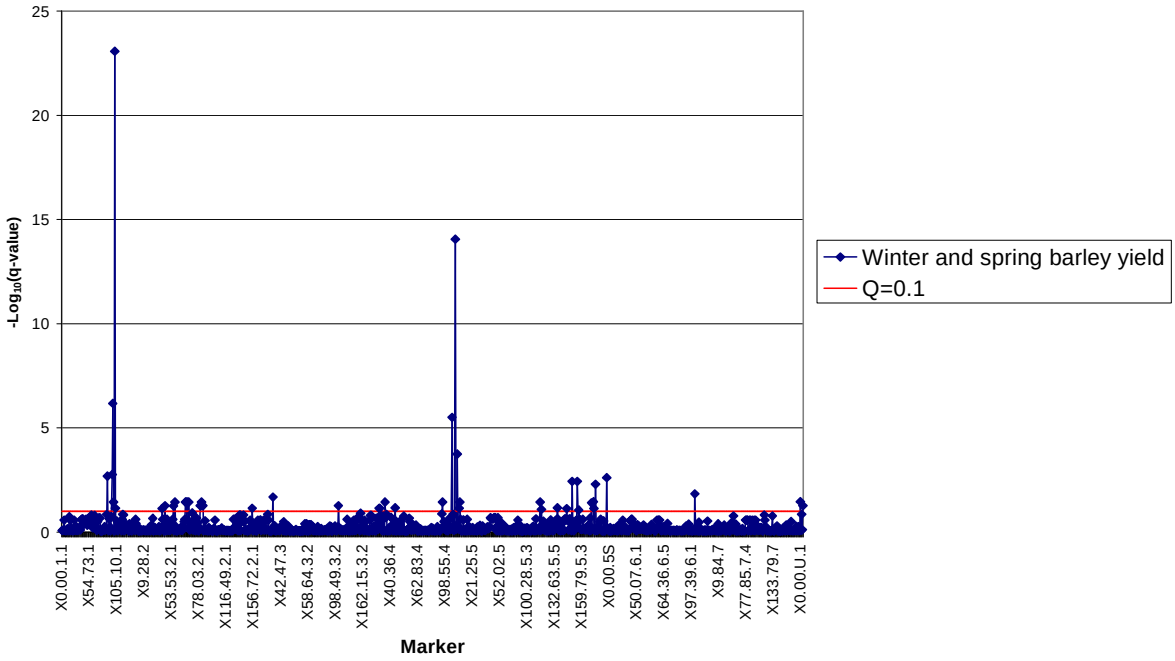


Figure 12-156. Association by MM. Markers arranged in genetic map order. Trait as indicated in legend.